Metagenomic Sequencing for Detection of Non-Traditional Biological Agents: Synthetic Constructs, Genetically Modified Organisms, and Emerging Threats

Project Report
for
Defence Research and Development Canada

Prepared by

Chronix Biomedical Inc.
5941 Optical Court, Suite 203E
San Jose, CA 95138

and the

Visual Genomics Centre
University of Calgary
Department of Biochemistry & Molecular Biology
Room 1150 Health Sciences Centre
3330 Hospital Drive NW
Calgary, AB T2N 4N1

PWGSC Contract Number: W7702-125259/001/EDM

Contract Scientific Authority: Chad Stratilo 403-544-4011 x4072

March 31, 2013

Contract Report
DRDC-RDDC-2014-C156

Table of contents

1 Introduction

1.1 Background

Non-traditional biological threat agents could potentially by-pass or defeat current detection technologies and thus it is important to address this gap. The most complete way to identify and characterize a traditional or non-traditional bacterial or viral agent is to sequence its entire genome. Not only does this identify the organism of interest, but the genome sequence can be used to also reveal targets for the management of the organism, such as putative antibiotic resistance (which may help guide treatment options); the presence of new virulence genes as new detection markers; and signs of genetic manipulation or engineering, including the insertion of synthetic biology constructs into the genome. This technology has the potential to mitigate credible white powder threat situations, where spore powders' are used but are in fact non-threatening insecticidal agents.

Recently, a new technology has entered the market (Nextra) that offers highly automated sample prep requiring only 15 minutes of hands-on time to produce a library for complete sequencing of a bacterial genome in less than eight hours using the bench top MiSeq next generation sequencer. This is much less than the time required to complete preliminary culturing on selective and differential media in traditional bacterial identification techniques. Sequencing technology has the potential to rapidly interrogate unknown sequences without prior knowledge and provide sequence data that can be used to distinguish credible threats from innocuous ones, inform treatment options, and reveal novel sequences of significance acquired through natural or engineered efforts.

With the acquisition of sequencing data becoming simpler and cheaper, the growing challenge is to develop technology especially the Bioinformatics pipelines that allow the sequence data to be translated into useful information. The specific Bioinformatics challenge lies in the rapid and accurate identification and detection of the gene or genes of interest within the massive amount of DNA sequence data generated.

1.2 Objectives

The goal of this project is to assess the feasibility of using present sequencing Bioinformatics technology for the identification of genetically modified organisms, identify any shortcomings and provide the input data for the development of a strategy for the ultimate discovery pipeline necessary for new biological agents.

2 Analysis of 16S rRNA gene sequences

2.1 Summary

Objective

It is known that 16S rRNA gene sequences can be used for taxonomic analysis of microbial communities. Using existing tools (Phoenix 2), currently in use for the discovery of novel species in the oil sands, we will assess the ability of this technology to quantify the relative abundance of specific microorganisms in a sample.

Conclusion

We were able to quantify relative abundance of organisms (at the genus level) in the 24 samples using 16S rRNA gene sequences. The 454 sequences have been analyzed by the Phoenix 2 pipeline. The

Illumina sequences have been analyzed using BLAST and the MEGAN metagenomic analysis tool. Both approaches succeeded in detecting most known organisms in a sample. For the samples with known relative abundance, we checked the accuracy of the identified relative abundance. In both approaches, there exists a wide range of similarity between the identified and known proportions. For example, *Klebsiella pneumoniae* was identified to comprise 55.66 % (known = 60 %) of Sample 1, with 454 sequences. *Yersinia tuberculosis* was identified to comprise 19.75 % (known = 20 %) of the same sample, with Illumina sequences. Identification of other organisms in the same sample does not necessarily produces this level of accuracy.

2.2 Grouping of samples

Among the 24 samples, 20 were partitioned into five groups (Groups 1, 2, 3, 4, and 5) to facilitate testing of organism identification performance, based on the given knowledge of known organisms in the samples. This means that each sample in a group is known to contain the same set of known organisms, but the proportions differ among the samples in the group. Four samples (Samples 5, 15, 16, and 22) do not belong to any group, as they are each known to contain a distinct set of organisms. The resulting groups and samples are described in the following table:.

Group or Sample ID	Organisms known?	Proportions of organisms known?	IDs of samples in group	IDs of organisms known to be contained
Group 1	Yes	Yes	1, 2, 3, 4	3, 11, 19, 27
Group 2	Yes	Yes	9, 10, 11, 12, 13	2, 12, 16, 18, 26, 27
Group 3	Yes	Yes	17, 18, 19, 20, 21	6, 7, 8, 9, 20, 21, 22, 23, 24, 29
Sample 22	Yes	Yes	22	8, 10, 11, 28
Sample 5	Yes	No	Pooled PCR from 1, 9, 11, 22	2, 3, 8, 10, 11, 12, 16, 18, 19, 26, 27, 28
Sample 15	Partially	No	Pooled PCR from 2, 12, 13, 14	2, 3, 11, 12, 16, 18, 19, 26, 27 and those in Sample 14
Sample 16	Partially	No	Pooled PCR from 10, 14, 18	2, 6, 7, 8, 9, 12, 16, 18, 20, 21, 22, 23, 24, 26, 27, 29 and those in Sample 14
Group 4	Yes	Yes (Sample 23) No (Sample 24)	23, 24	2, 3, 6, 7, 8, 9, 10, 11, 12, 16, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29
Group 5	No	No	6, 7, 8, 14	None (environmental samples)

The organisms known to be contained in each group or sample are as follows:

- Group 1:
  - *Bacillus thuringiensis*, *Klebsiella pneumoniae*, *Vibrio Harvey*, and *Yersinia pseudotuberculosis* (4 organisms)
- Group 2:
  - *Escherichia coli*, *Yersinia enterocolitica*, *Yersinia pseudotuberculosis*, *Mycobacterium tuberculosis*, *Streptococcus pneumoniae*, and *Staphylococcus aureus* (6 organisms)
- Group 3:

- *Acinetobacter baumannii*, *Neisseria meningitides*, *Klebsiella pneumoniae*, *Salmonella enterica*, *Rhizobium radiobacter*, *Clostridium difficile*, *Staphylococcus aureus*, *Ochrobactrum anthropui*, *Legionella penumonphia*, and *Shigella dysteria* (10 organisms)

- Sample 22: *Klebsiella pneumonia* with four different strains
- Sample 5:
  - *Escherichia coli*, *Bacillus thuringiensis*, *Klebsiella pneumoniae*, *Yersinia enterocolitica*, *Yersinia pseudotuberculosis*, *Mycobacterium tuberculosis*, *Streptococcus pneumoniae*, *Vibrio harveyi*, *Staphylococcus aureus* (9 organisms)
- Sample 15:
  - *Escherichia coli*, *Bacillus thuringiensis*, *Klebsiella pneumoniae*, *Yersinia enterocolitica*, *Yersinia pseudotuberculosis*, *Mycobacterium tuberculosis*, *Streptococcus pneumoniae*, *Vibrio harveyi*, and *Staphylococcus aureus* (9 organisms) plus unknown organisms in Sample 14
- Sample 16:
  - *Escherichia coli*, *Acinetobacter baumannii*, *Neisseria meningitidis*, *Klebsiella pneumoniae*, *Salmonella enterica*, *Yersinia enterocolitica*, *Yersinia pseudotuberculosis, Mycobacterium tuberculosis, Streptococcus pneumoniae*, *Rhizobium radiobacter*, *Clostridium difficile*, *Staphylococcus aureus*, *Ochrobactrum anthropui*, *Legionella pneumonphia*, and *Shigella dysteria* (15 organisms) plus unknown organisms in Sample 14
- Group 4:
  - *Escherichia coli*, *Bacillus thuringiensis*, *Acinetobacter baumannii*, *Neisseria meningitides*, *Salmonella enterica, Klebsiella pneumoniae*, *Yersinia enterocolitica*, *Yersinia pseudotuberculosis*, *Mycobacterium tuberculosis*, *Streptococcus pneumoniae*, *Vibrio harveyi*, *Rhizobium radiobacter*, *Clostridium difficile, Staphylococcus aureus*, *Ochrobactrum anthropui*, *Legionella pneumonphia*, and *Shigella dysteria* (17 organisms)
- Group 5: Unknown organisms

2.3 Processing of 454 sequences

Each sample or group of samples was submitted to Phoenix 2 16S rRNA analysis pipeline hosted by the Visual Genomics Centre at the University of Calgary. Each analysis job submission resulted in multiple sets of taxonomic classification results, depending on two factors: clustering distance cutoff used for clustering reads into operational taxonomic units (OTUs) and the classification algorithm.

For the analysis of **relative abundance of organisms**, we have used the RDP classifier with SILVA training dataset using the OTUs formed at the clustering distance of 0.05. Both the clustering of reads and classification OTUs were done using the mothur package, which is incorporated in the Phoenix 2 pipeline. Taxa have been assigned to OTUs down to the genus level, which is the lowest classification resolution provided by the SILVA training dataset.

2.4 Processing of Illumina sequences

We first checked the overall quality of the Illumina 16S rRNA samples using the FastQC sequence quality checking tool. Given a FASTQ file as input, this tool provides a set of statistical measurements and their visualizations. It was discovered that the quality of the reads from the R2 files of the paired-end sequencing was consistently poor, to such a degree that it would not be worthwhile to use them for

assembly. For example, for most R2 reads, the base quality scores were around 2 and the base call was mostly N after the base number position 40. This made the read lengths too short after trimming low-quality bases. Therefore, only the R1 files were used for further processing.

Based on the same FastQC quality pre-check, the quality of raw reads from each sample was checked as follows:

- Trim from the end until the quality score of the last base is at least 20
- Filter if the read length (after trimming) is less than 75, there is an ambiguous (N) base, or the average quality score over all bases is less than 25

For the assembly of the short reads, we used the Velvet assembler. After experimenting with a few different hash lengths for the assembly, we chose 71 as the *k*-mer length to use. The results of the quality control using the above parameters and the assembly using Velvet are summarized in the table below.

Sample ID	**Quality control**		**Assembly using Velvet** (Hash length 71)					
	Raw reads	**Good reads**	**Median coverage depth**	**Number of all contigs**	**N50**	**Max contig length**	**Total assembly length**	**Number of high-coverage contigs used for analysis**
1	31,030,170	28,302,874 (91.2%)	3.24	156,950	106	867	2,655,289	368
2	29,788,587	25,772,592 (86.5%)	3.00	183,857	105	625	2,766,285	366
3	28,769,609	24,904,421 (86.6%)	2.94	166,001	104	607	2,424,049	391
4	31,490,206	27,997,249 (88.9%)	2.96	192,383	104	606	2,774,047	340
5	32,121,143	28,416,384 (88.5%)	2.87	230,818	104	1,234	3,233,942	633
6	33,842,963	30,086,287 (88.9%)	3.21	331,567	105	650	4,658,446	4,261
7	33,434,347	29,196,116 (87.3%)	3.90	296,607	108	5,643	4,727,024	2,600
8	28,320,782	25,052,158 (88.5%)	3.40	284,614	105	681	4,116,568	4,173
9	27,444,176	23,495,837 (85.6%)	3.08	168,550	107	837	2,770,642	431
10	34,615,575	29,973,821 (86.6%)	3.00	171,324	106	808	2,711,403	535
11	30,191,652	26,952,255 (89.3%)	2.93	173,614	106	992	2,695,567	337
12	33,714,019	29,670,792 (88.0%)	3.00	213,578	106	832	3,396,452	447
13	28,254,744	25,418,082 (90.0%)	2.72	189,956	105	855	2,771,555	990
14	33,357,964	29,647,193	3.68	371,242	106	588	5,359,676	3,816

		(88.9%)						
15	33,202,583	28,532,192 (85.9%)	2.94	234,608	104	864	3,220,235	715
16	32,288,970	28,402,038 (88.0%)	2.76	233,401	104	724	3,307,855	1,801
17	28,312,124	23,845,221 (84.2%)	2.84	166,237	104	4,078	2,292,034	633
18	33,138,907	25,413,342 (76.7%)	3.23	188,706	109	505	2,826,567	691
19	28,278,215	24,014,335 (84.9%)	2.83	140,519	107	867	2,175,052	719
20	27,535,077	23,053,100 (83.7%)	2.77	119,181	108	1,219	1,957,047	633
21	30,061,189	25,158,834 (83.7%)	3.00	203,086	105	882	2,924,326	760
22	31,631,710	26,689,105 (84.4%)	3.69	143,188	110	657	2,680,381	238
23	31,370,341	26,887,721 (85.7%)	2.95	252,493	100	841	3,026,857	820
24	36,231,171	29,959,540 (82.7%)	2.80	254,686	102	552	3,264,906	998

For each sample, a set of high-coverage contigs was used for organism identification. The Velvet default threshold of half the median coverage depth was used for selecting the contigs to analyze. The number of such contigs for each sample is shown in the last column of the above table.

For the analysis of **relative abundance of organisms**, BLAST (blastn) was run on the selected contigs for each sample against the NCBI "nt" database. Then we used MEGAN to parse the BLAST output and run the taxonomy analysis. The proportions of the identified taxa, which are to be compared with the known proportions, were calculated not by the number of contigs assigned to the taxon, but by the sum of the median coverages of all the contigs assigned to the taxon. This way, the number of reads assembled into a contig is considered when we estimate the abundance of the organism to which the contig has been assigned.

2.5 Identification of organisms and estimation of their proportions

For each 16S rRNA sample, we have performed two separate analyses based on 454 sequences and Illumina sequences, respectively, to identify organisms present in the sample and to estimate their relative abundance. The analysis results for each sample are presented side by side per organism, to facilitate the comparison of the differences between the 454-based and Illumina-based approaches.

For each sample, a table shows the identified organisms from each analysis and their estimated proportions. In the table, **red boldface** letters and percentages in the "Species known" column indicate the known organisms and their proportions. The **boldface** letters and percentages in the "Genus identified" column indicate the same organisms identified by the two (454 and Illumina) analyses, at the genus level. If a table cell in the identification column is empty, it indicates that the particular genus was not identified by the analysis. After this table, a taxonomy tree from the Illumina sequence analysis of the sample is shown, expanded down to the genus level.

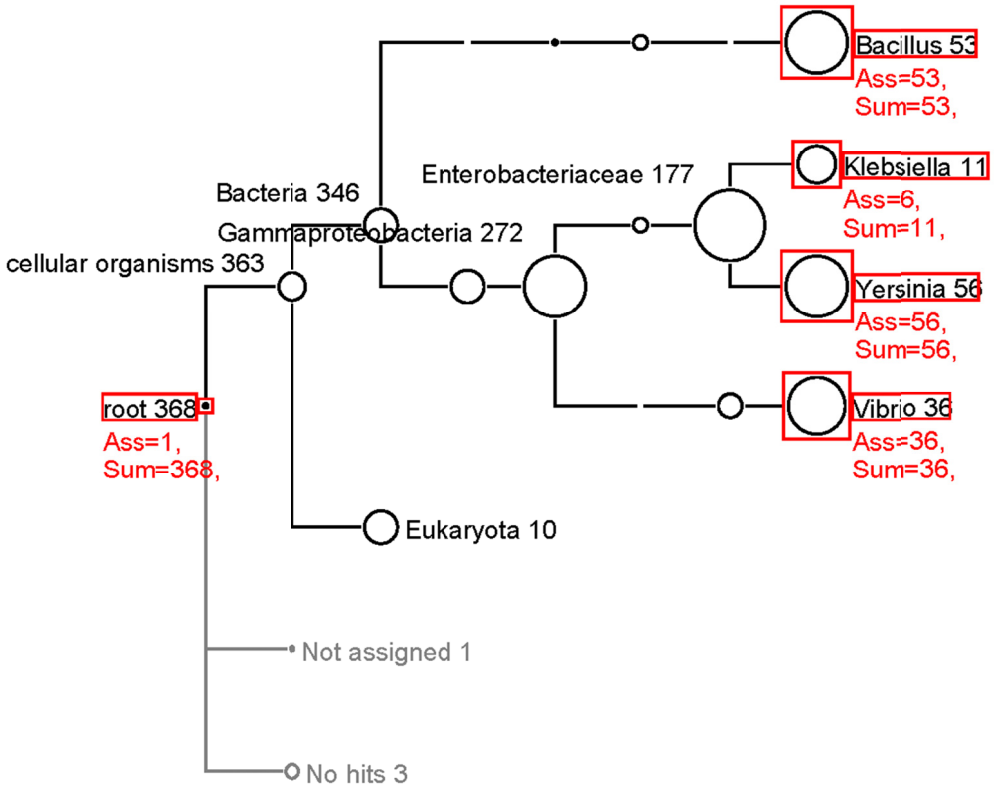
2.5.1 Group 1

Group 1 consists of Samples 1, 2, 3, and 4.

Organisms identified in Sample 1

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Klebsiella pneumoniae (11)	**60**	**Klebsiella**	**55.66**	**7.63**
Yersinia pseudotuberculosis (27)	**20**	**Yersinia**	**10.59**	**19.75**
Bacillus thuringiensis (3)	**15**	**Bacillus**	**28.21**	**16.03**
Vibrio harvey (19)	**5**	**Vibrio**	**3.17**	**7.33**
		Enterobacter	0.60	
		Kurthia	0.30	

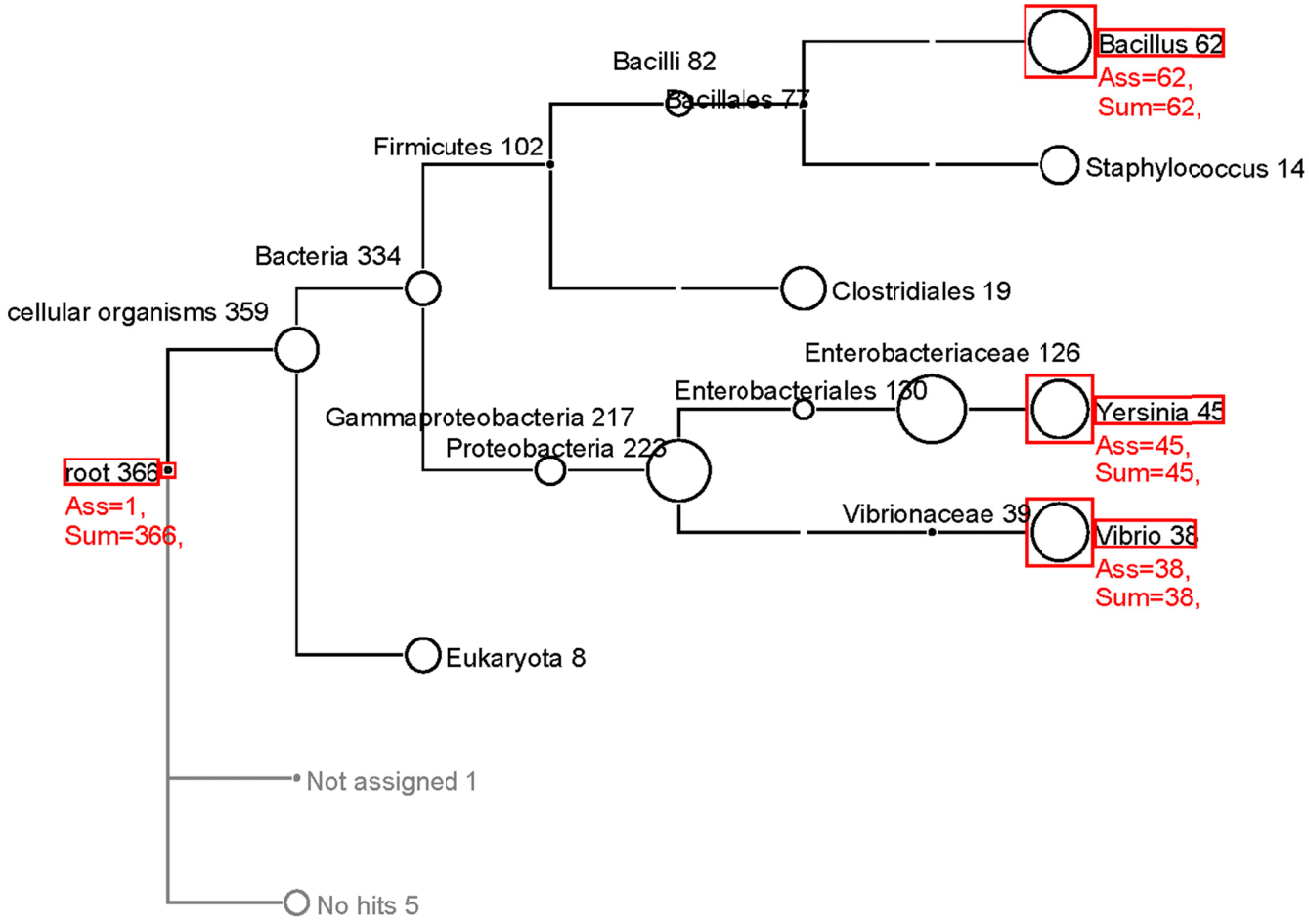
Genus-level taxonomy tree from Illumina sequence analysis



Organisms identified in Sample 2

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Bacillus thuringiensis (3)	**30**	**Bacillus**	**42.80**	**16.02**
Klebsiella pneumoniae (11)	**30**	**Klebsiella**	**29.65**	
Yersinia pseudotuberculosis (27)	**30**	**Yersinia**	**17.00**	**18.52**
Vibrio Harvey (19)	**10**	**Vibrio**	**7.47**	**12.29**
		Staphylococcus		1.23
		Enterobacter	0.58	
		Kurthia	0.39	
		Saccharomyces	0.23	
		Staphylococcus	0.08	
		Acinetobacter	0.04	

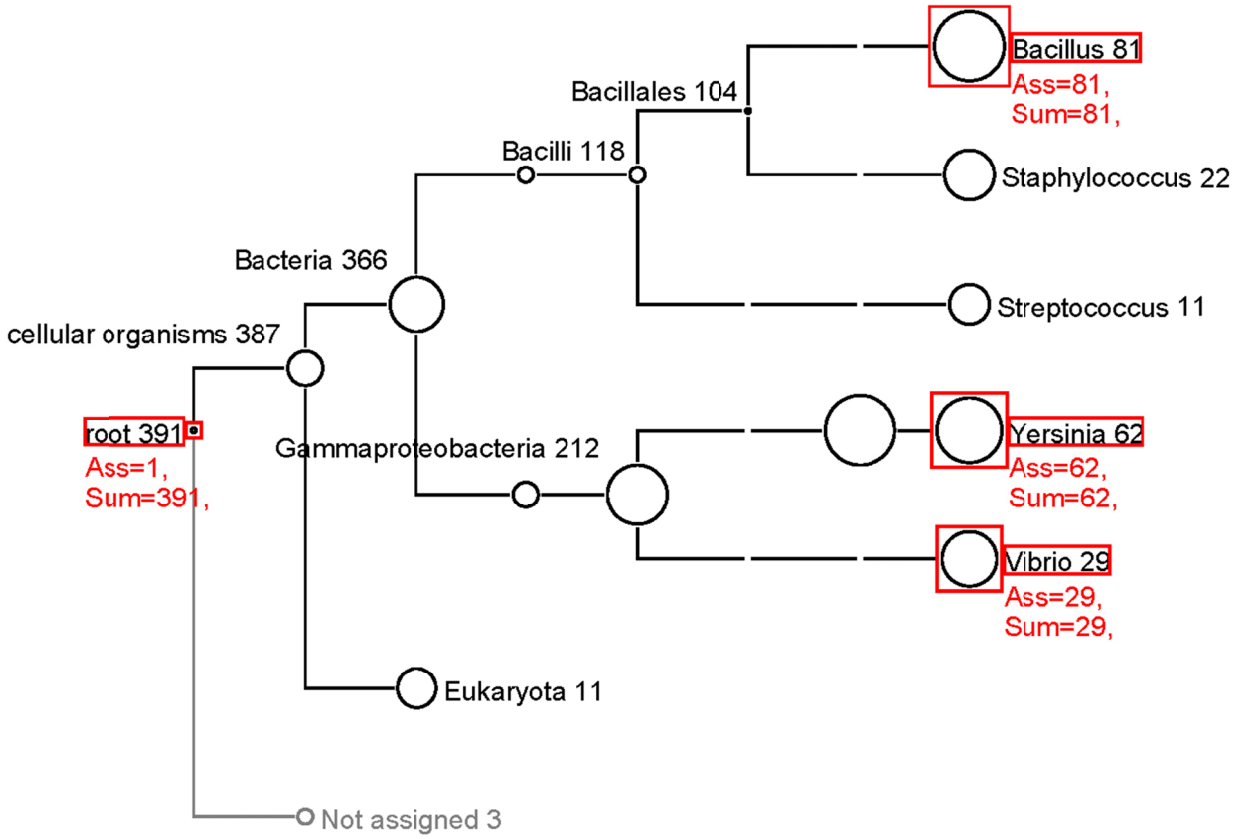
Genus-level taxonomy tree from Illumina sequence analysis of Sample 2



Organisms identified in Sample 3

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Yersinia pseudotuberculosis (27)	**60**	**Yersinia**	**38.82**	**14.22**
Klebsiella pneumoniae (11)	**20**	**Klebsiella**	**23.68**	
Bacillus thuringiensis (3)	**10**	**Bacillus**	**24.34**	**22.54**
Vibrio Harvey (19)	**10**	**Vibrio**	**8.99**	**9.47**
		Staphylococcus		1.81
		Streptococcus		0.96
		Enterobacter	0.44	

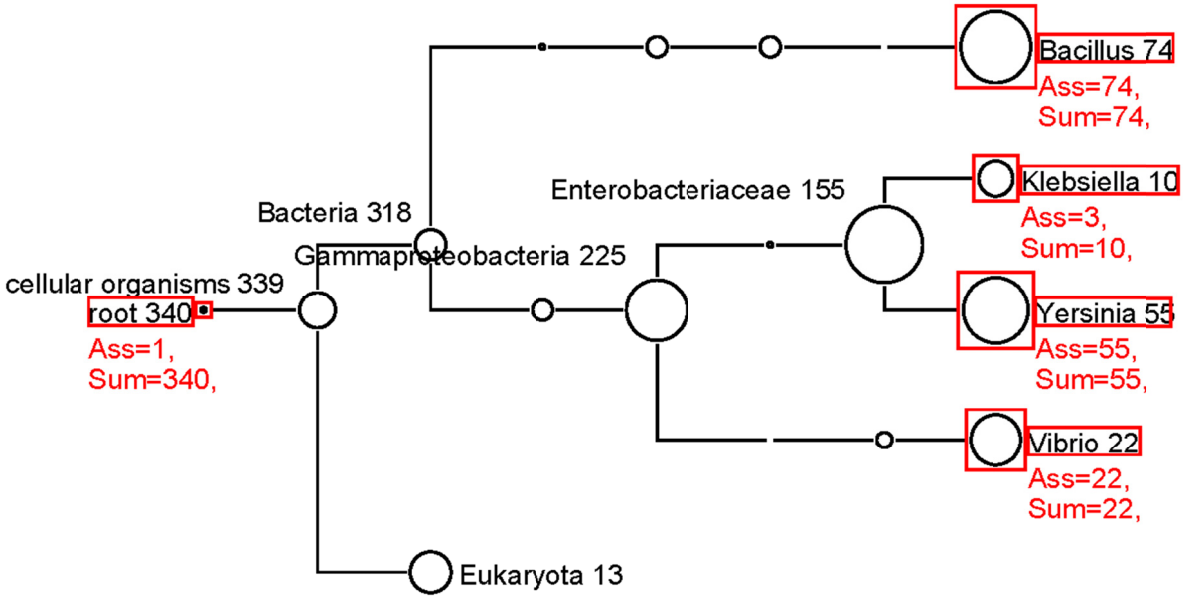
Genus-level taxonomy tree from Illumina sequence analysis of Sample 3



Organisms identified in Sample 4

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Klebsiella pneumoniae (11)	**40**	**Klebsiella**	**43.30**	**14.49**
Yersinia pseudotuberculosis (27)	**40**	**Yersinia**	**24.21**	**13.95**
Bacillus thuringiensis (3)	**10**	**Bacillus**	**20.35**	**21.58**
Vibrio Harvey (19)	**10**	**Vibrio**	**8.88**	**6.77**
		Enterobacter	0.55	
		Kurthia	0.22	
		Saccharomyces	0.06	
		Acinetobacter	0.06	
		Thermoplasma	0.06	

Genus-level taxonomy tree from Illumina sequence analysis of Sample 4



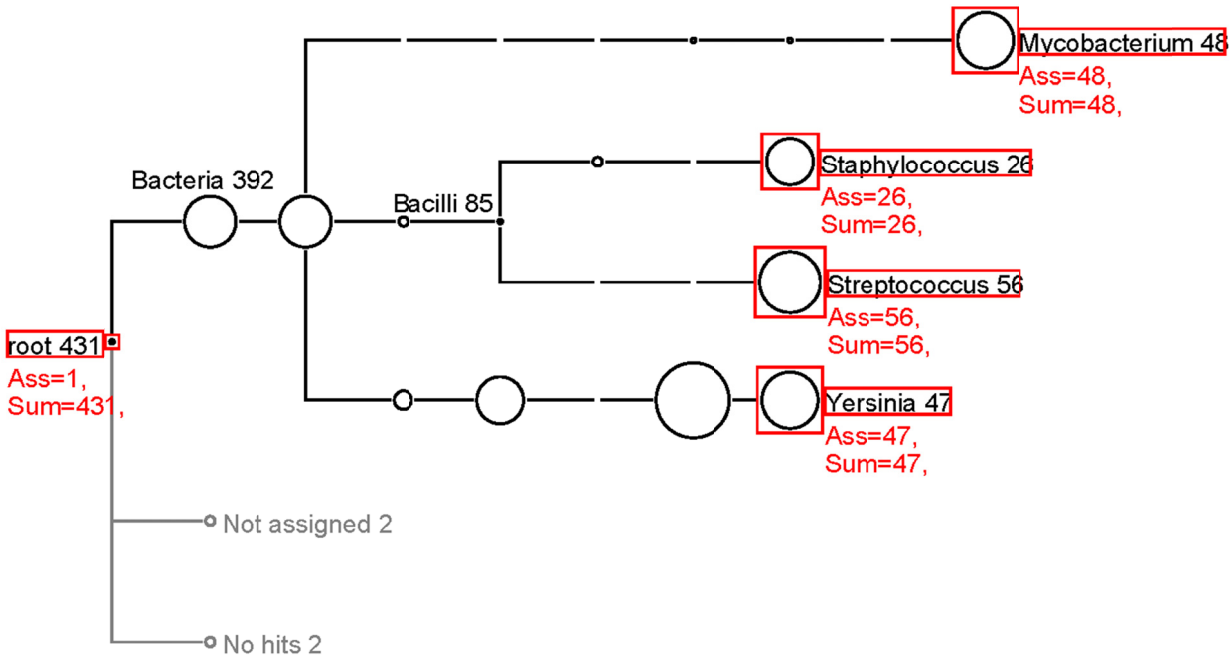
2.5.2 Group 2

Group 2 consists of Samples 9, 10, 11, 12, and 13.

Organisms identified in Sample 9

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Yersinia enterocolitica (12)	**25**	**Yersinia**	**0.30**	**10.87**
Yersinia pseudotuberculosis (27)	**10**			
Mycobacterium tuberculosis (16)	**25**	**Mycobacterium**	**0.55**	**7.20**
Streptococcus pneumoniae (18)	**10**	**Streptococcus**	**33.10**	**4.27**
Escherichia coli (2)	**10**	**Escherichia**	**4.93**	
Staphylococcus aureus (26)	**10**	**Staphylococcus**	**12.17**	**5.05**
		Cronobacter	0.20	

Genus-level taxonomy tree from Illumina sequence analysis of Sample 9

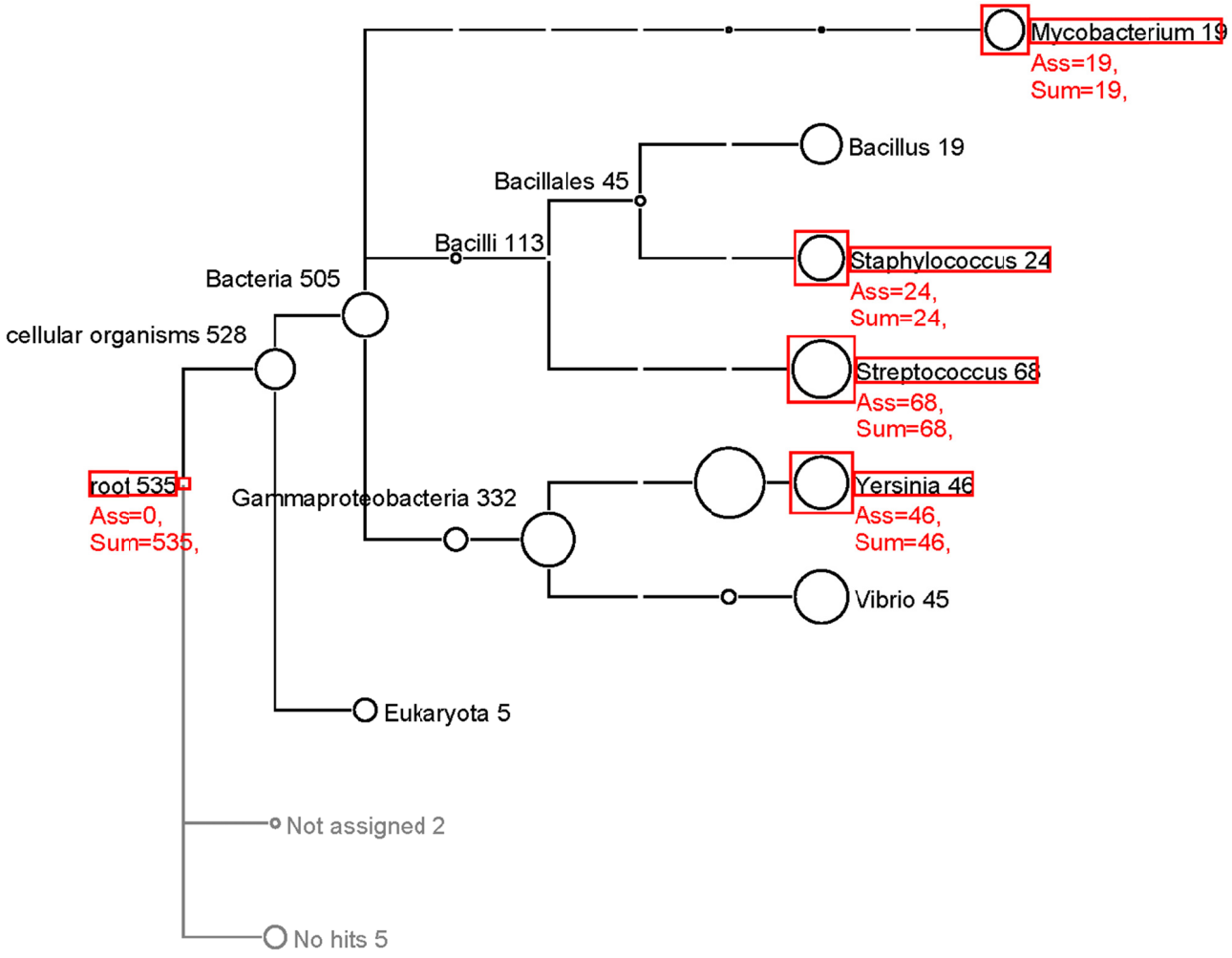


Organisms identified in Sample 10

Species known		Genus identified		
Species (ID)	**%***	**Genus**	**454 %**	**Illumina %**
Yersinia enterocolitica (12)	**25**	**Yersinia**	**0.48**	**4.01**
Yersinia pseudotuberculosis (27)	**20**			
Mycobacterium tuberculosis (16)	**15**	**Mycobacterium**	**0.24**	**1.62**
Streptococcus pneumoniae (18)	**15**	**Streptococcus**	**26.56**	**5.80**
Escherichia coli (2)	**10**	**Escherichia**	**3.74**	
Staphylococcus aureus (26)	**10**	**Staphylococcus**	**22.74**	**2.08**
		Vibrio		2.16
		Bacillus		1.13

* Sum of known percentages = 105 %

Genus-level taxonomy tree from Illumina sequence analysis of Sample 10

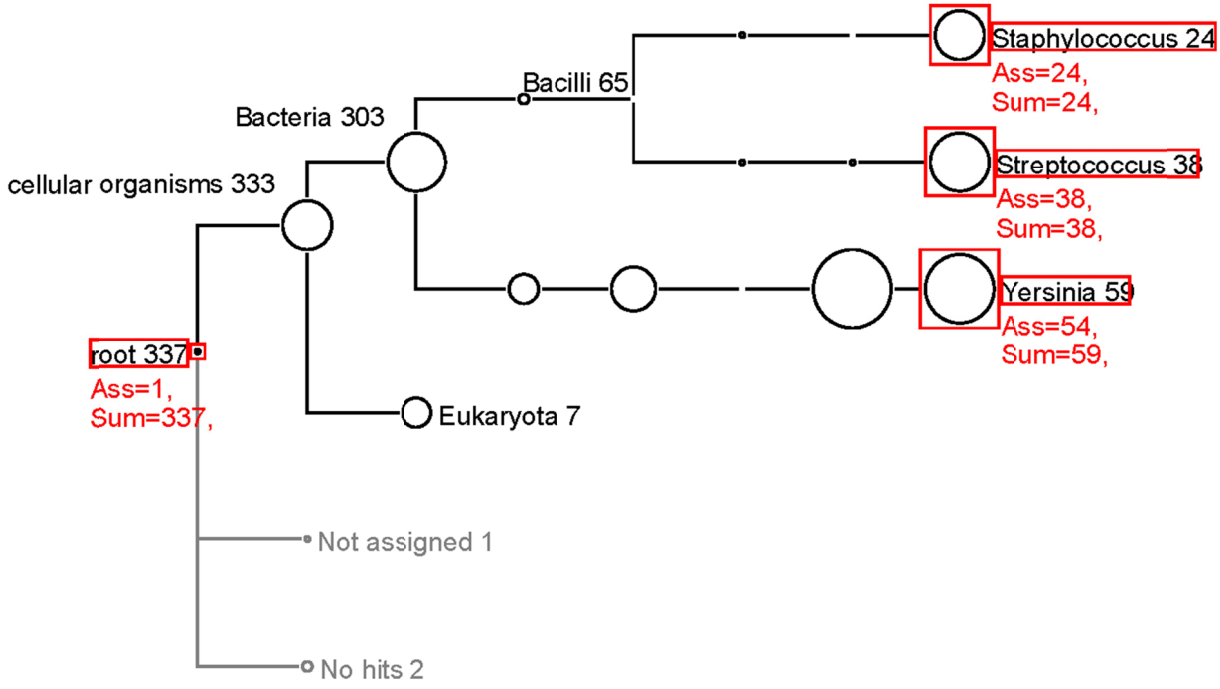


Organisms identified in Sample 11

Species known		Genus identified		
Species (ID)	**%***	**Genus**	**454 %**	**Illumina %**
Yersinia enterocolitica (12)	**35**	**Yersinia**	**0.39**	**9.39**
Yersinia pseudotuberculosis (27)	**30**			
Streptococcus pneumoniae (18)	**20**	**Streptococcus**	**23.84**	**9.82**
Escherichia coli (2)	**10**	**Escherichia**	**3.61**	
Staphylococcus aureus (26)	**10**	**Staphylococcus**	**20.58**	**4.02**
Mycobacterium tuberculosis (16)	**5**	**Mycobacterium**	**0.04**	

* Sum of known percentages = 110 %

Genus-level taxonomy tree from Illumina sequence analysis of Sample 11

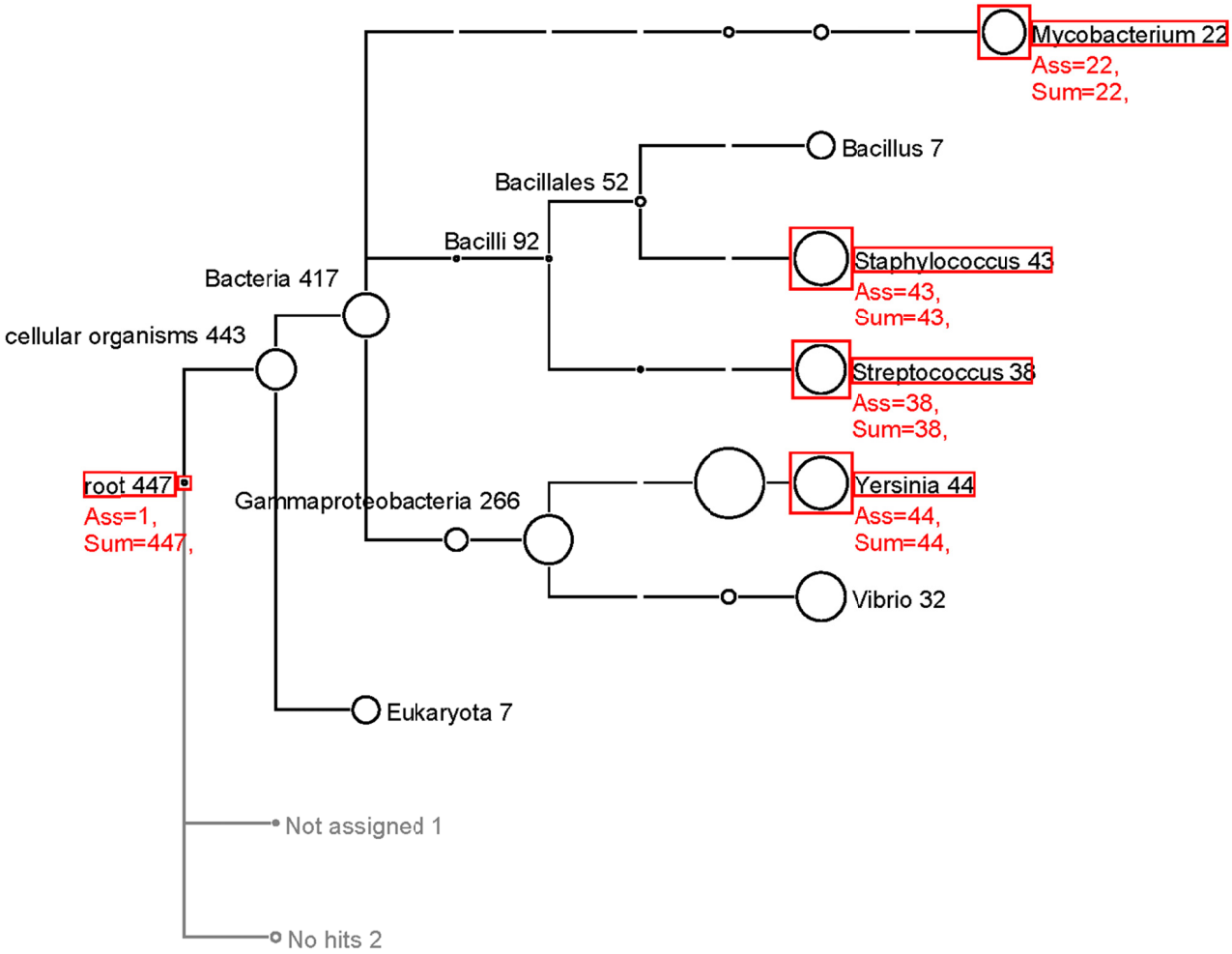


Organisms identified in Sample 12

Species known		Genus identified		
Species (ID)	**%***	**Genus**	**454 %**	**Illumina %**
Yersinia enterocolitica (12)	**25**	**Yersinia**	**0.39**	**7.34**
Yersinia pseudotuberculosis (27)	**20**			
Streptococcus pneumoniae (18)	**20**	**Streptococcus**	**24.99**	**5.78**
Escherichia coli (2)	**15**	**Escherichia**	**5.84**	
Mycobacterium tuberculosis (16)	**5**	**Mycobacterium**	**0.20**	**3.26**
Staphylococcus aureus (26)	**5**	**Staphylococcus**	**29.85**	**10.94**
		Vibrio		3.10
		Bacillus		0.86

* Sum of known percentages = 90 %

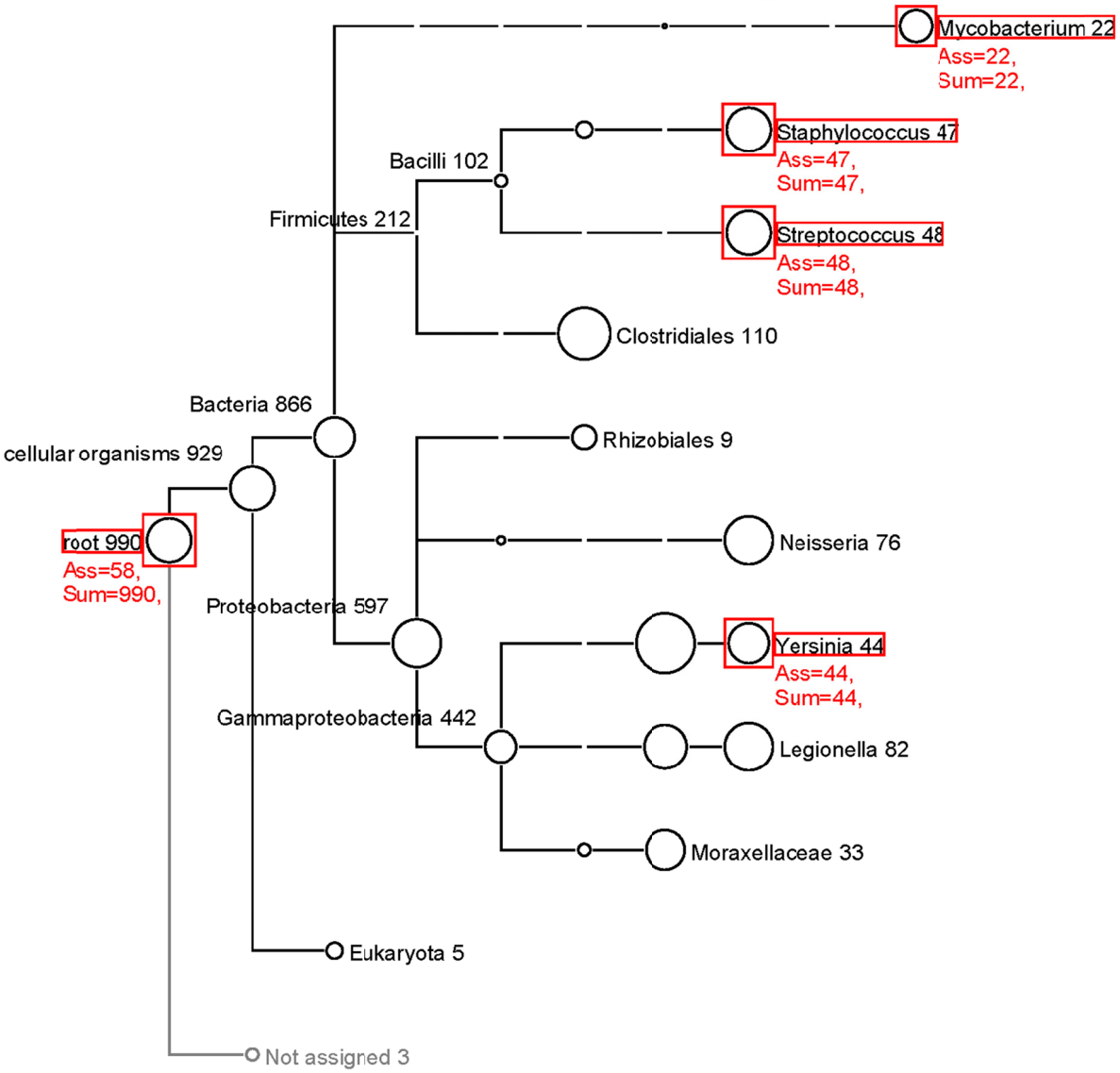
Genus-level taxonomy tree from Illumina sequence analysis of Sample 12



Organisms identified in Sample 13

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Yersinia enterocolitica (12)	**25**	**Yersinia**	**0.28**	**4.21**
Yersinia pseudotuberculosis (27)	**10**			
Streptococcus pneumoniae (18)	**25**	**Streptococcus**	**18.97**	**5.70**
Staphylococcus aureus (26)	**20**	**Staphylococcus**	**27.74**	**3.28**
Escherichia coli (2)	**10**	**Escherichia**	**4.08**	
Mycobacterium tuberculosis (16)	**10**	**Mycobacterium**	**0.10**	**1.20**
		Legionella	2.23	7.32
		Neisseria	0.05	5.22
		Ochrobactrum	0.28	
		Rhizobium	0.10	
		Klebsiella	0.05	

Genus-level taxonomy tree from Illumina sequence analysis of Sample 13



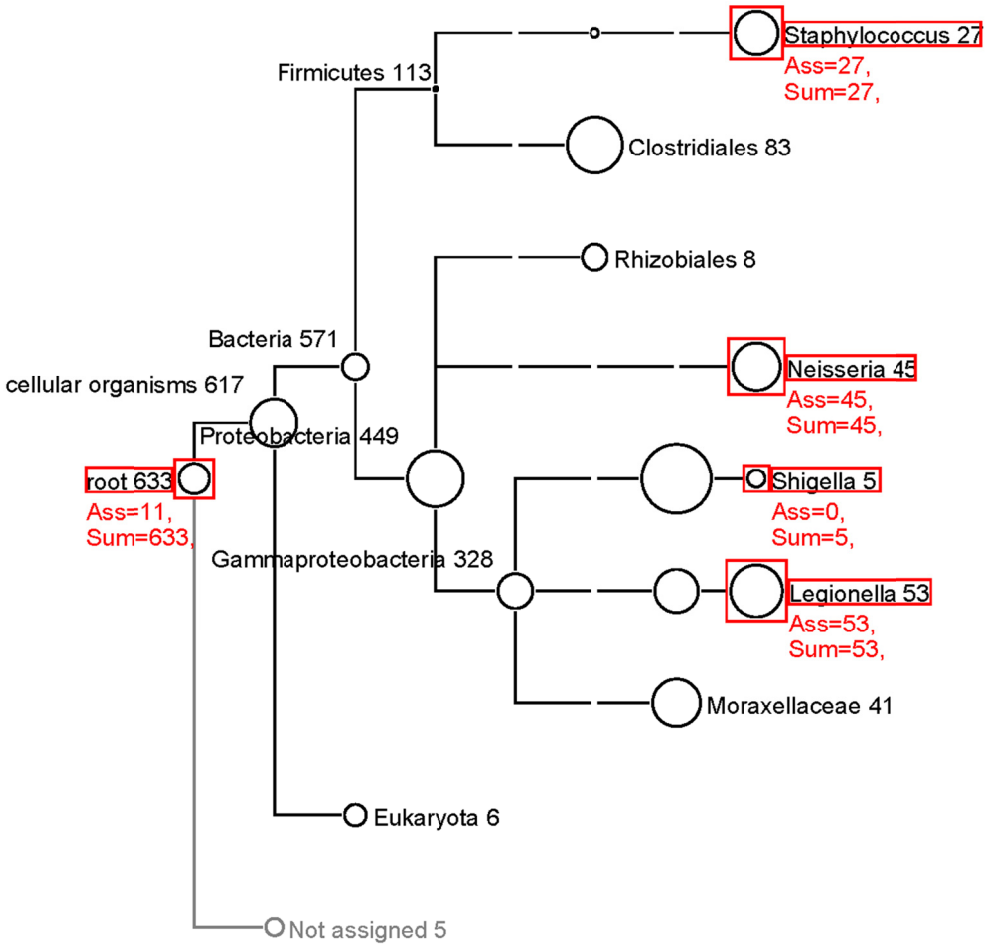
2.5.3 Group 3

Group 3 consists of Samples 17, 18, 19, 20, and 21.

Organisms identified in Sample 17

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Acinetobacter baumannii (6)	**10**	**Acinetobacter**	**11.45**	
Neisseria meningitides (7)	**10**	**Neisseria**	**0.16**	**4.28**
Klebsiella pneumoniae (8)	**10**	**Klebsiella**		
Salmonella enterica (9)	**10**	**Salmonella**		
Rhizobium radiobacter (20)	**10**	**Rhizobium**	**0.16**	
Clostridium difficile (21)	**10**	**Clostridium**		
Staphylococcus aureus (22)	**10**	**Staphylococcus**	**17.02**	**3.07**
Ochrobactrum anthropui (23)	**10**	**Ochrobactrum**	**1.85**	
Legionella penumonphia (24)	**10**	**Legionella**	**3.75**	**9.78**
Shigella dysteria (29)	**10**	**Shigella**		**0.41**
		Escherichia	3.34	

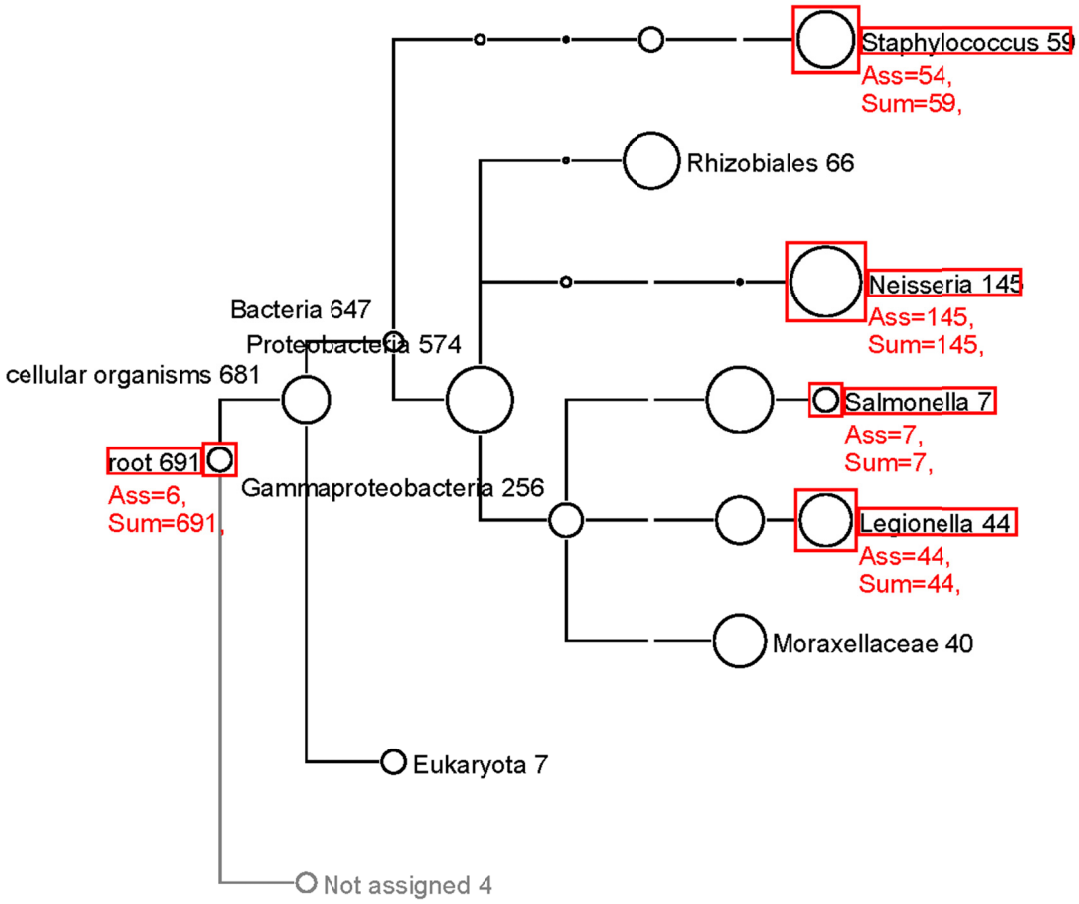
Genus-level taxonomy tree from Illumina sequence analysis of Sample 17



Organisms identified in Sample 18

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Acinetobacter baumannii (6)	**15**	**Acinetobacter**	**33.26**	
Salmonella enterica (9)	**15**	**Salmonella**		**0.09**
Rhizobium radiobacter (20)	**15**	**Rhizobium**	**1.09**	
Staphylococcus aureus (22)	**15**	**Staphylococcus**	**47.19**	**0.81**
Legionella penumonphia (24)	**15**	**Legionella**	**9.44**	**2.54**
Neisseria meningitides (7)	**5**	**Neisseria**	**0.64**	**3.63**
Klebsiella pneumoniae (8)	**5**	**Klebsiella**	**0.05**	
Clostridium difficile (21)	**5**	**Clostridium**		
Ochrobactrum anthropui (23)	**5**	**Ochrobactrum**	**1.91**	
Shigella dysteria (29)	**5**	**Shigella**		
		Escherichia	4.13	

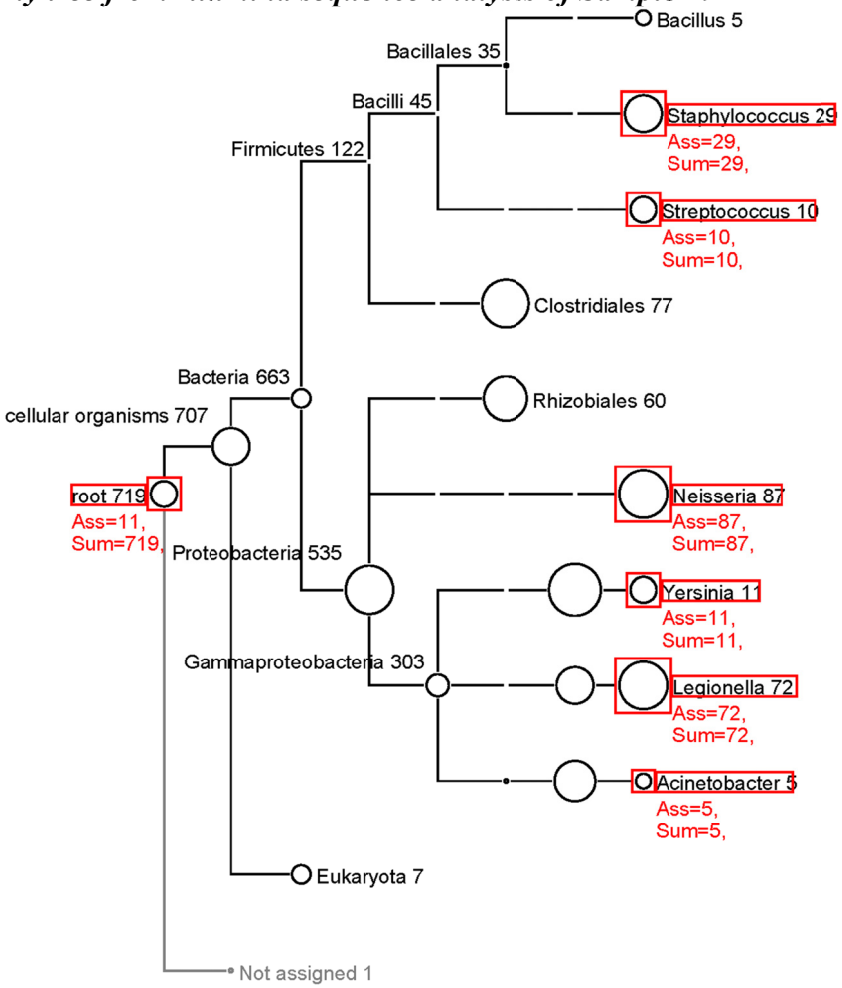
Genus-level taxonomy tree from Illumina sequence analysis of Sample 18



Organisms identified in Sample 19

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Neisseria meningitides (7)	**15**	**Neisseria**	**1.06**	**12.64**
Klebsiella pneumoniae (8)	**15**	**Klebsiella**	**0.04**	
Clostridium difficile (21)	**15**	**Clostridium**		
Ochrobactrum anthropui (23)	**15**	**Ochrobactrum**	**4.60**	
Shigella dysteria (29)	**15**	**Shigella**		
Acinetobacter baumannii (6)	**5**	**Acinetobacter**	**6.21**	**0.67**
Salmonella enterica (9)	**5**	**Salmonella**		
Rhizobium radiobacter (20)	**5**	**Rhizobium**	**0.12**	
Staphylococcus aureus (22)	**5**	**Staphylococcus**	**6.48**	**3.24**
Legionella penumonphia (24)	**5**	**Legionella**	**2.16**	**8.48**
		Escherichia	5.77	
		Bacillus		1.79
		Yersinia		0.40
		Peptostreptococcus	0.16	
		Cronobacter	0.04	

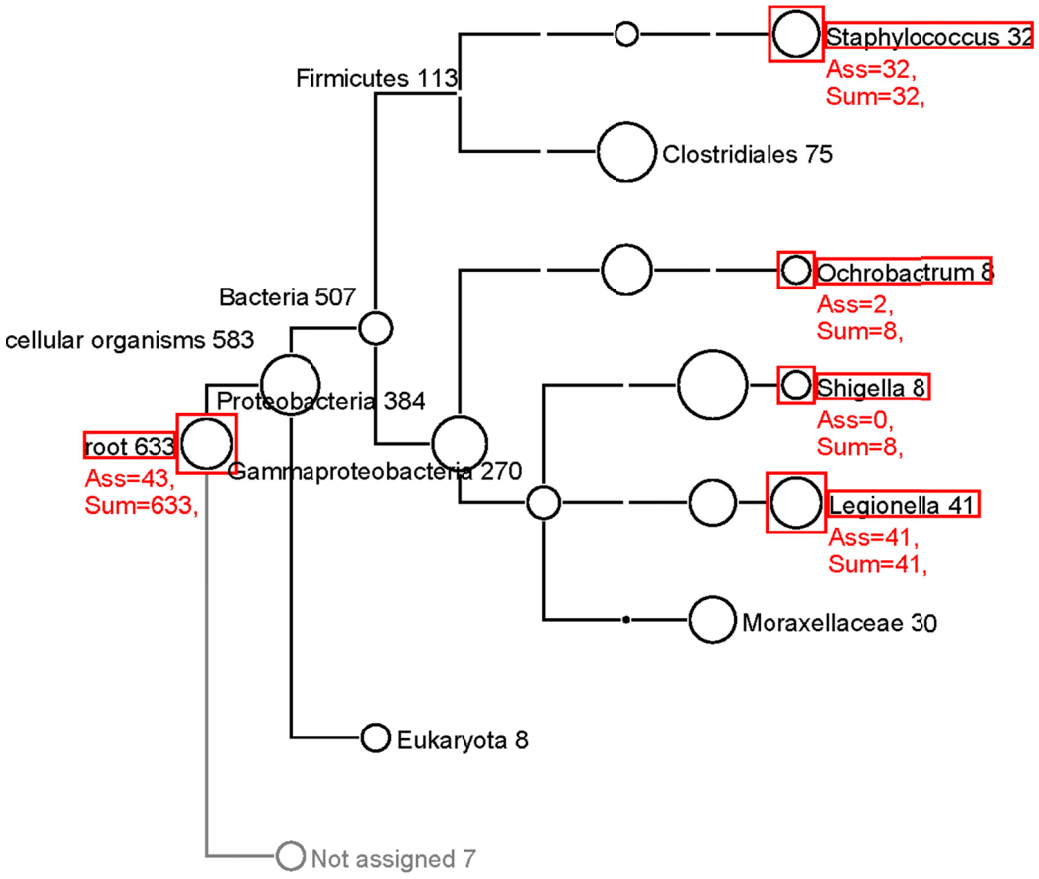
Genus-level taxonomy tree from Illumina sequence analysis of Sample 19



Organisms identified in Sample 20

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Rhizobium radiobacter (20)	**19**	**Rhizobium**	**0.39**	
Clostridium difficile (21)	**19**	**Clostridium**		
Staphylococcus aureus (22)	**19**	**Staphylococcus**	**29.87**	**1.88**
Ochrobactrum anthropui (23)	**19**	**Ochrobactrum**	**4.00**	**0.38**
Legionella penumonphia (24)	**19**	**Legionella**	**9.65**	**8.93**
Acinetobacter baumannii (6)	**1**	**Acinetobacter**	**0.70**	
Neisseria meningitides (7)	**1**	**Neisseria**		
Klebsiella pneumoniae (8)	**1**	**Klebsiella**		
Salmonella enterica (9)	**1**	**Salmonella**		
Shigella dysteria (29)	**1**	**Shigella**		**0.70**
		Escherichia	0.70	
		Brucella	0.04	

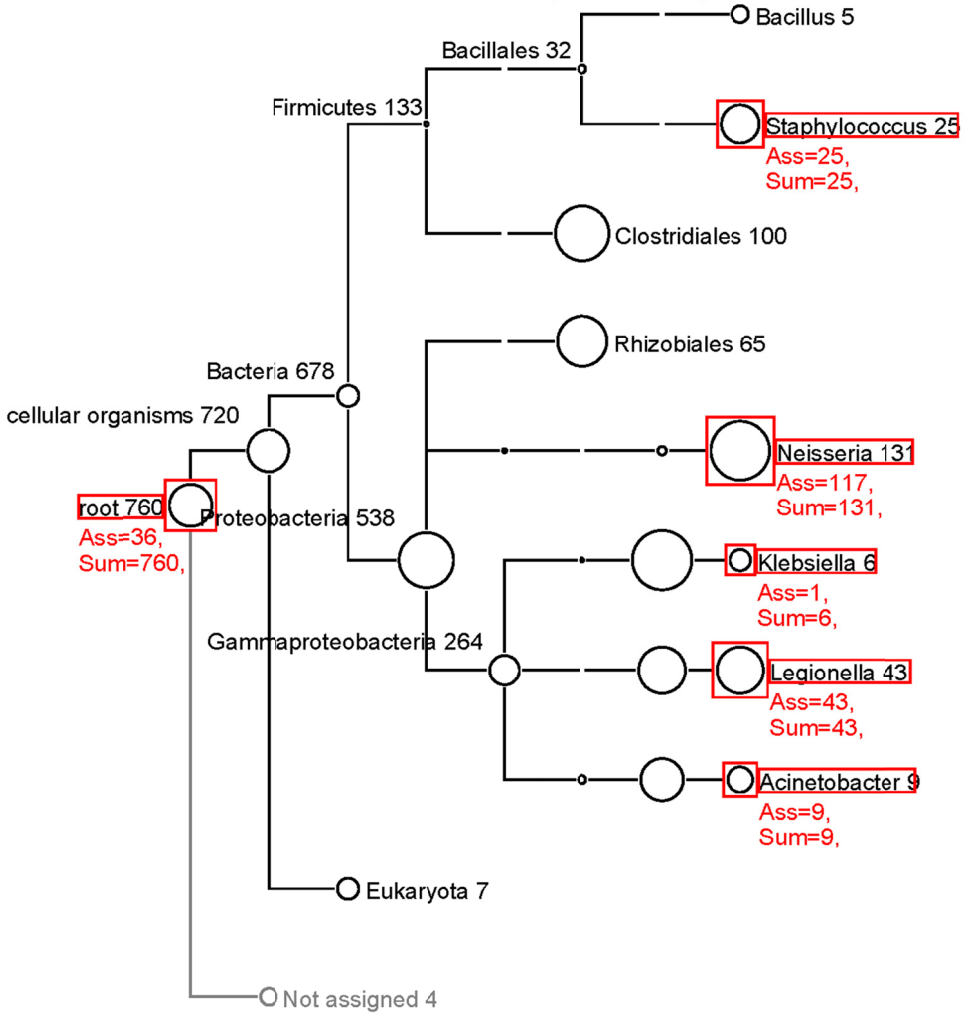
Genus-level taxonomy tree from Illumina sequence analysis of Sample 20



Organisms identified in Sample 21

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Acinetobacter baumannii (6)	**17**	**Acinetobacter**	**33.92**	
Neisseria meningitides (7)	**17**	**Neisseria**	**1.35**	**19.95**
Klebsiella pneumoniae (8)	**17**	**Klebsiella**	**0.06**	**0.27**
Salmonella enterica (9)	**17**	**Salmonella**		
Shigella dysteria (29)	**17**	**Shigella**		
Rhizobium radiobacter (20)	**3**	**Rhizobium**	**0.18**	
Clostridium difficile (21)	**3**	**Clostridium**		
Staphylococcus aureus (22)	**3**	**Staphylococcus**	**16.69**	**2.89**
Ochrobactrum anthropui (23)	**3**	**Ochrobactrum**	**2.51**	
Legionella penumonphia (24)	**3**	**Legionella**	**1.41**	**3.85**
		Escherichia	10.45	
		Bacillus		0.98
		Moraxella	0.12	
		Bergeriella	0.06	

Genus-level taxonomy tree from Illumina sequence analysis of Sample 21

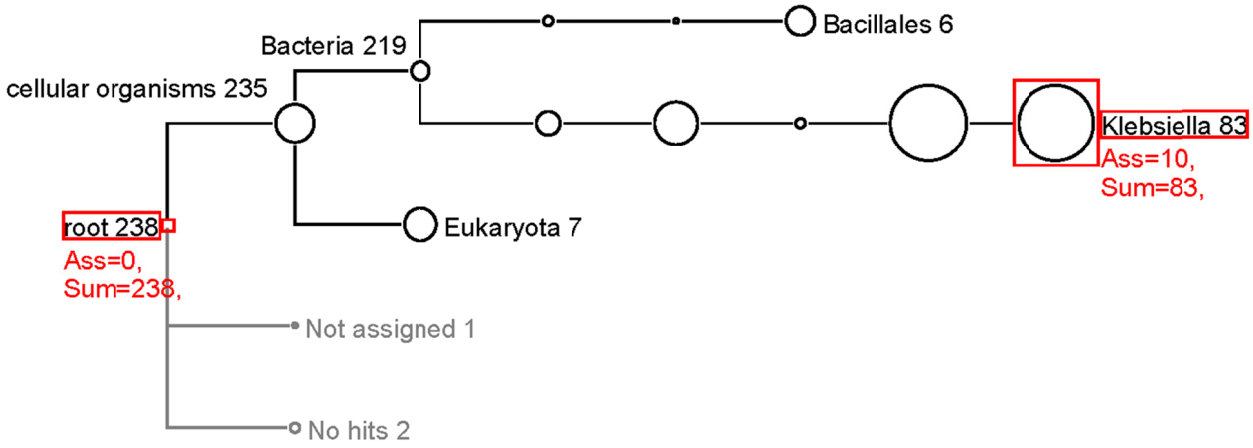


2.5.4 Sample 22

Organisms identified in Sample 22

Species known		Genus identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**
Klebsiella pneumoniae (28)	**40**	**Klebsiella**	**99.56**	**15.28**
Klebsiella pneumoniae (8)	**20**			
Klebsiella pneumoniae (10)	**20**			
Klebsiella pneumoniae (11)	**20**			

Genus-level taxonomy tree from Illumina sequence analysis of Sample 22



2.5.5 Sample 5

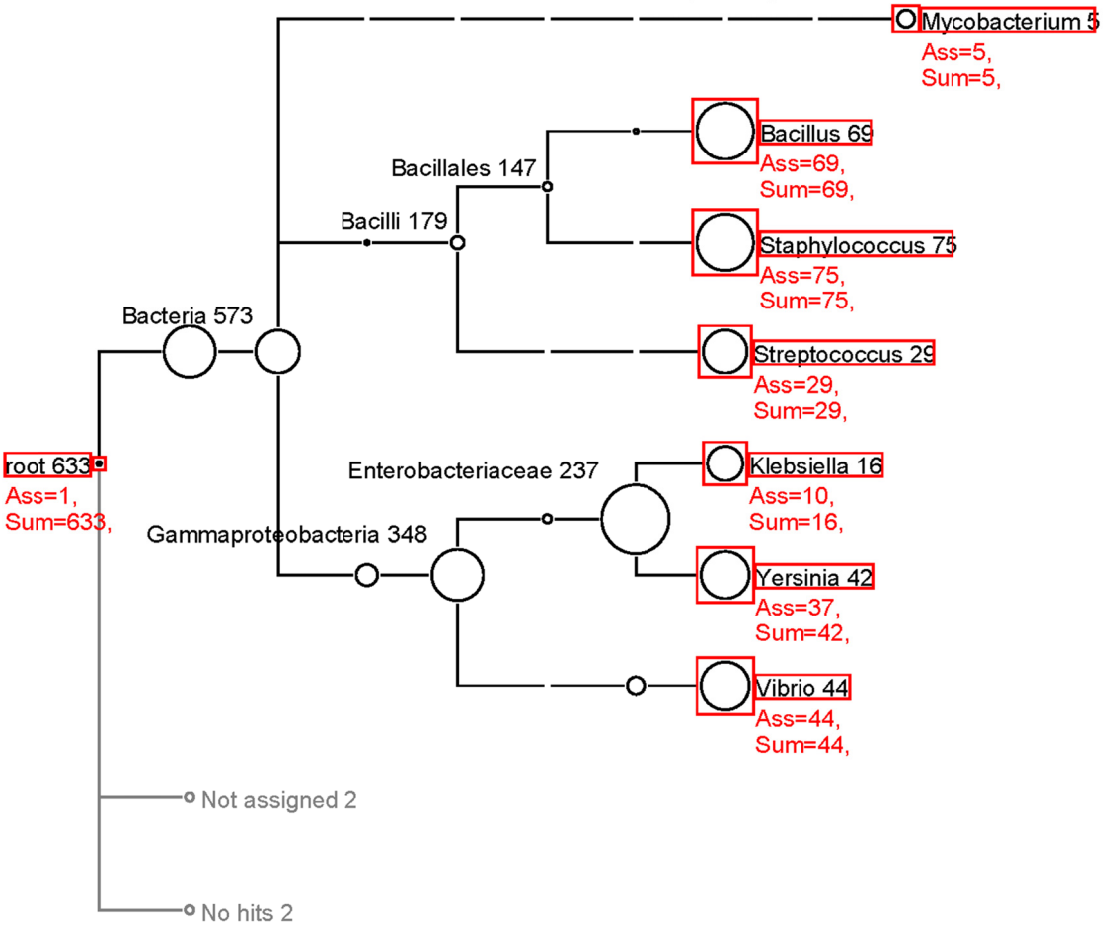
Sample 5 contains pooled PCR products from Samples 1, 9, 11, and 22.

.

Organisms identified in Sample 5

Species known (ID)	Genus identified		
	Genus	454 %	Illumina %
Escherichia coli (2)	**Escherichia**	**0.16**	
Bacillus thuringiensis (3)	**Bacillus**	**10.83**	**14.42**
Klebsiella pneumoniae (8, 10, 11, 28)	**Klebsiella**	**43.63**	**0.72**
Yersinia enterocolitica (12)	**Yersinia**	**21.16**	**5.18**
Yersinia pseudotuberculosis (27)			
Mycobacterium tuberculosis (16)	**Mycobacterium**	**0.08**	**0.38**
Streptococcus pneumoniae (18)	**Streptococcus**	**15.26**	**5.84**
Vibrio harveyi (19)	**Vibrio**	**2.28**	**4.41**
Staphylococcus aureus (26)	**Staphylococcus**	**4.93**	**13.75**
	Raoultella	0.24	
	Cronobacter	0.08	

Genus-level taxonomy tree from Illumina sequence analysis of Sample 5



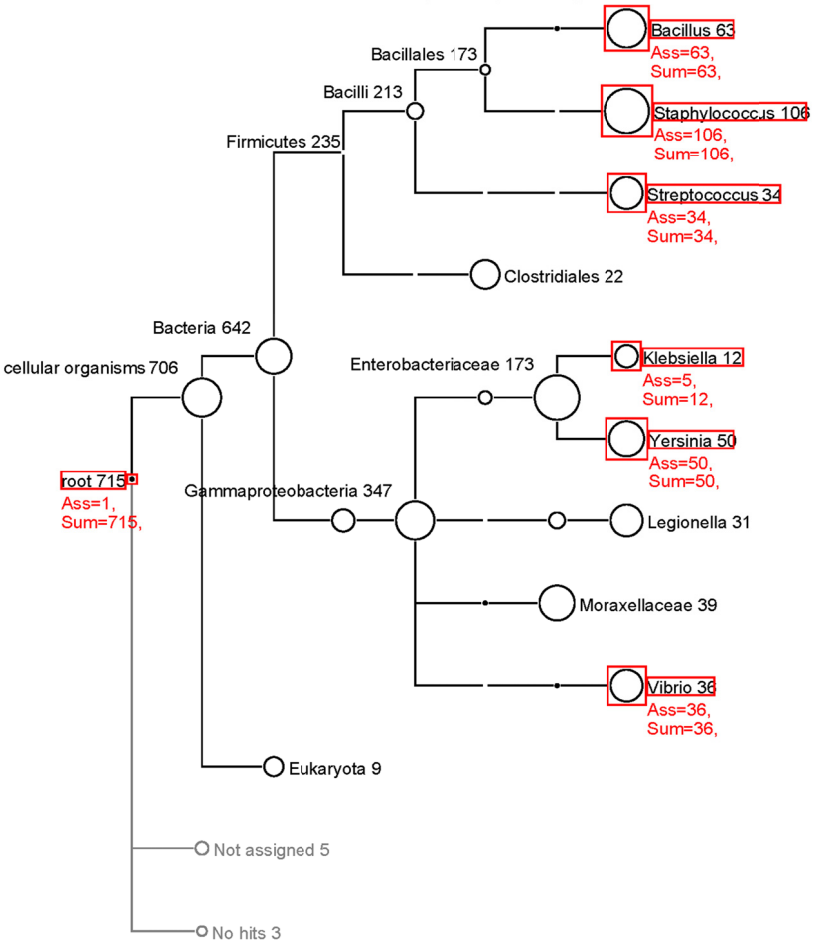
2.5.6 Sample 15

Sample 15 contains pooled PCR products from Samples 2, 12, 13, and 14.

Organisms identified in Sample 15

Species known (ID)	Genus identified		
	Genus	454 %	Illumina %
Escherichia coli (2)	**Escherichia**	**0.96**	
Bacillus thuringiensis (3)	**Bacillus**	**23.24**	**9.68**
Klebsiella pneumoniae (11)	**Klebsiella**	**1.12**	**6.01**
Yersinia enterocolitica (12)	**Yersinia**	**21.01**	**4.78**
Yersinia pseudotuberculosis (27)			
Mycobacterium tuberculosis (16)	**Mycobacterium**		
Streptococcus pneumoniae (18)	**Streptococcus**	**4.63**	**4.96**
Vibrio harveyi (19)	**Vibrio**	**6.31**	**4.43**
Staphylococcus aureus (26)	**Staphylococcus**	**4.55**	**18.18**
	Legionella		1.38
	Acinetobacter	0.40	
	Raoultella	0.40	
	Legionella	0.08	

Genus-level taxonomy tree from Illumina sequence analysis of Sample 15



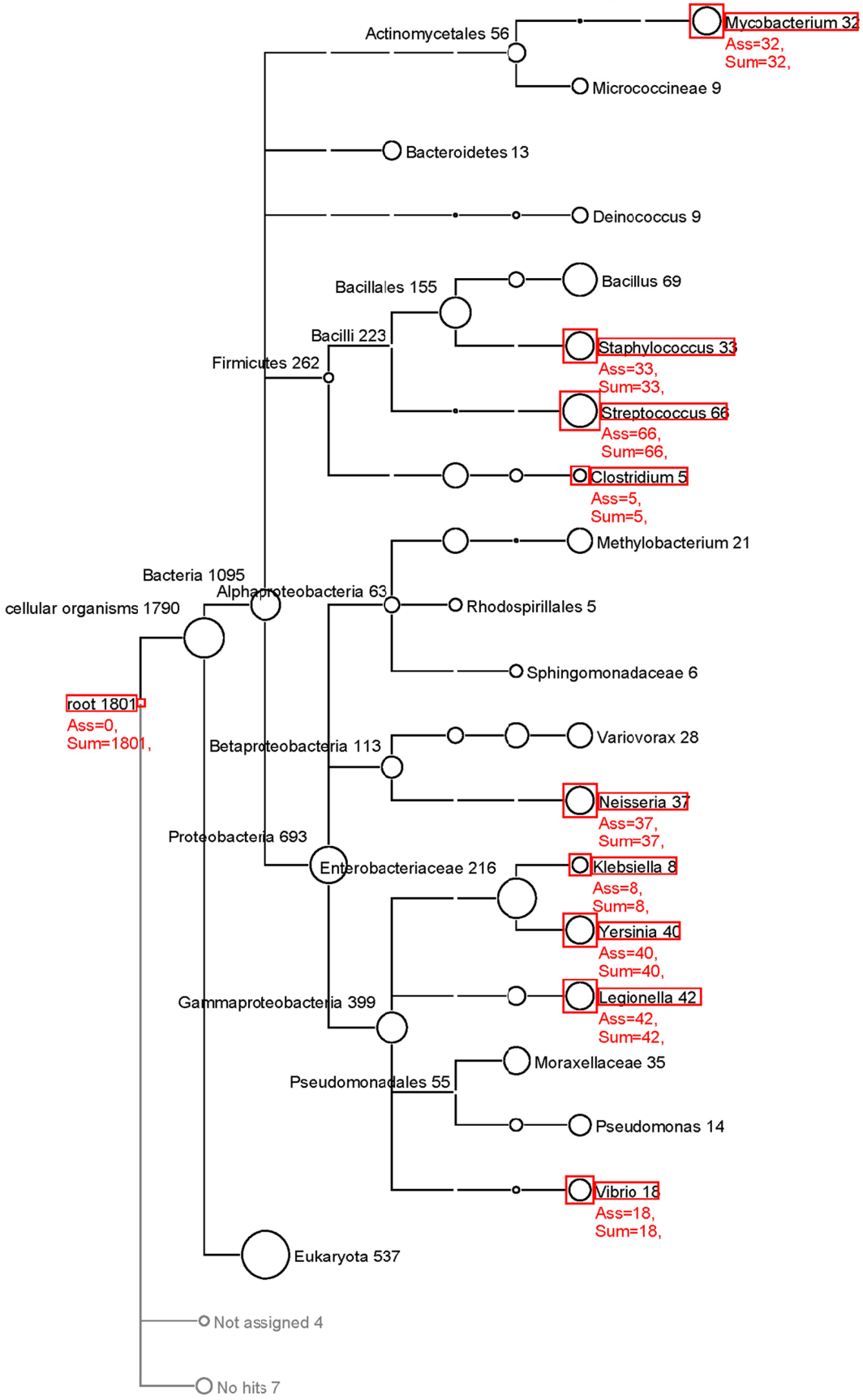
2.5.7 Sample 16

Sample 16 contains pooled PCR products from Samples 10, 14, and 18.

Organisms identified in Sample 16

Species known (ID)	Genus identified		
	Genus	454 %	Illumina %
Escherichia coli (2)	**Escherichia**	**0.10**	
Acinetobacter baumannii (6)	**Acinetobacter**	**18.60**	
Neisseria meningitidis (7)	**Neisseria**	**0.46**	**1.10**
Klebsiella pneumoniae (8)	**Klebsiella**	**17.47**	**0.17**
Salmonella enterica (9)	**Salmonella**		
Yersinia enterocolitica (12) **Yersinia pseudotuberculosis** (27)	**Yersinia**	**0.51**	**2.96**
Mycobacterium tuberculosis (16)	**Mycobacterium**	**0.21**	**0.80**
Streptococcus pneumoniae (18)	**Streptococcus**	**6.15**	**5.81**
Rhizobium radiobacter (20)	**Rhizobium**	**0.72**	
Clostridium difficile (21)	**Clostridium**		**0.44**
Staphylococcus aureus (22, 26)	**Staphylococcus**	**21.82**	**0.83**
Ochrobactrum anthropui (23)	**Ochrobactrum**	**0.31**	
Legionella pneumonphia (24)	**Legionella**	**3.43**	**3.06**
Shigella dysteria (29)	**Shigella**		
	Bacillus	0.67	3.52
	Variovorax		1.07
	Pseudomonas	0.92	0.43
	Massilia	0.87	
	Methylobacterium	0.26	0.82
	Vibrio	0.05	0.79
	Columnosphaeria	0.67	
	Herbaspirillum	0.67	
	Deinococcus		0.57
	Acidovorax	0.51	
	Sporidiobolus	0.36	
	Kirschsteiniothelia	0.26	
	Hericium	0.26	
	Sphingomonas	0.21	
	Herminiimonas	0.15	
	Paenibacillus	0.15	
	Arthrobacter	0.10	
	Bergeriella	0.10	
	Buchnera	0.10	
	Brevundimonas	0.05	
	Bensingtonia	0.05	
	Duganella	0.05	
	Filobasidium	0.05	
	Naxibacter	0.05	
	Mesorhizobium	0.05	

Genus-level taxonomy tree from Illumina sequence analysis of Sample 16



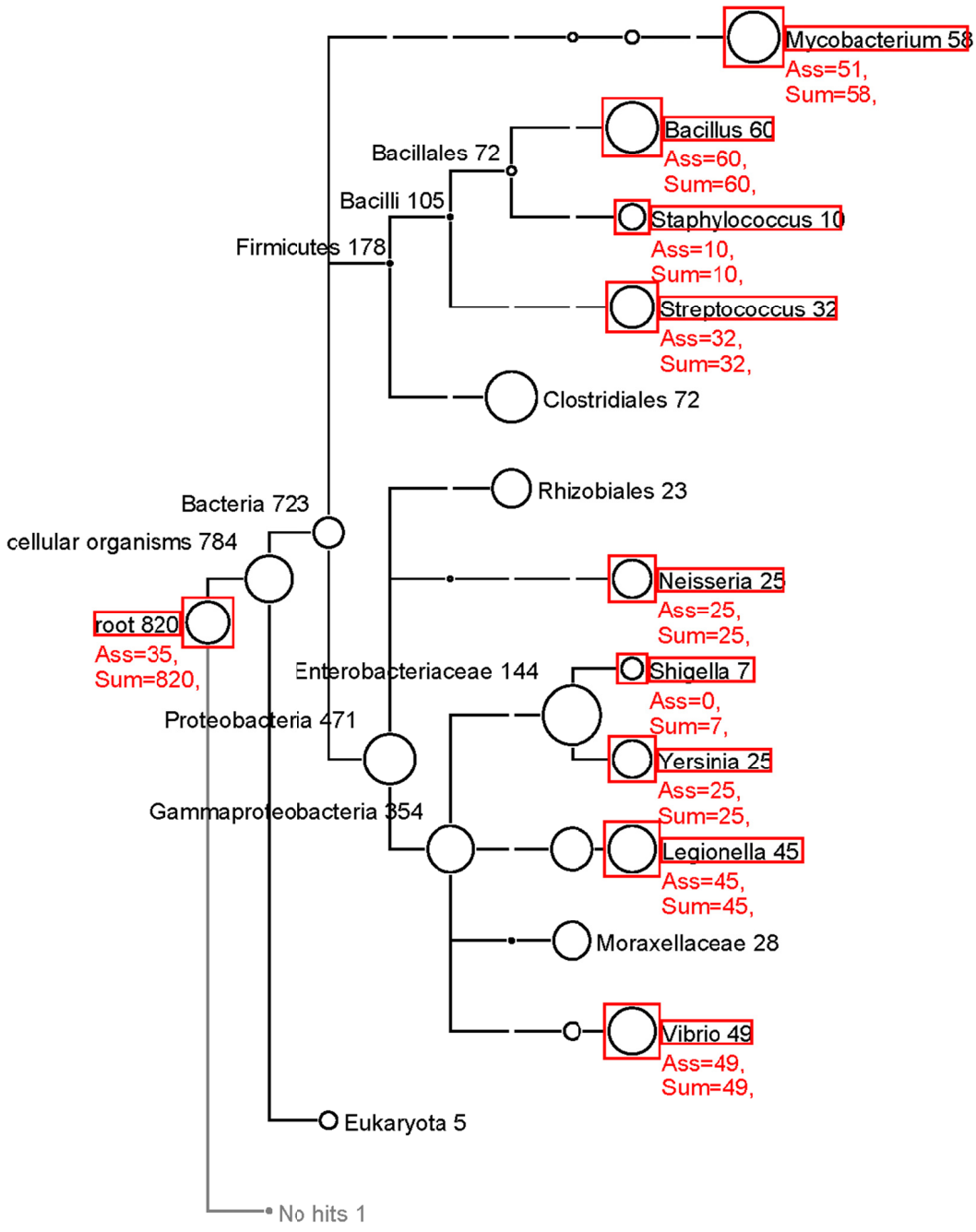
2.5.8 Group 4

Group 4 consists of Samples 23 and 24.

Organisms identified in Sample 23

Species known (ID)	Genus identified		
	Genus	454 %	Illumina %
Escherichia coli (2)	**Escherichia**	**0.55**	
Bacillus thuringiensis (3)	**Bacillus**	**2.00**	**7.42**
Acinetobacter baumannii (6)	**Acinetobacter**	**4.56**	
Neisseria meningitides (7)	**Neisseria**	**0.95**	
Klebsiella pneumoniae (8, 10, 11, 28)	**Klebsiella**	**13.21**	**3.02**
Salmonella enterica (9)	**Salmonella**		
Yersinia enterocolitica (12)	**Yersinia**	**8.61**	**2.70**
Yersinia pseudotuberculosis (27)			
Mycobacterium tuberculosis (16)	**Mycobacterium**	**0.55**	**6.22**
Streptococcus pneumoniae (18)	**Streptococcus**	**5.01**	**3.33**
Vibrio harveyi (19)	**Vibrio**	**3.35**	**5.61**
Rhizobium radiobacter (20)	**Rhizobium**	**0.75**	
Clostridium difficile (21)	**Clostridium**		
Staphylococcus aureus (22, 26)	**Staphylococcus**	**10.01**	**0.67**
Ochrobactrum anthropui (23)	**Ochrobactrum**	**1.05**	
Legionella pneumonphia (24)	**Legionella**	**4.35**	**5.25**
Shigella dysteria (29)	**Shigella**		**0.87**
	Peptostreptococcus	0.10	
	Raoultella	0.05	

Genus-level taxonomy tree from Illumina sequence analysis of Sample 23

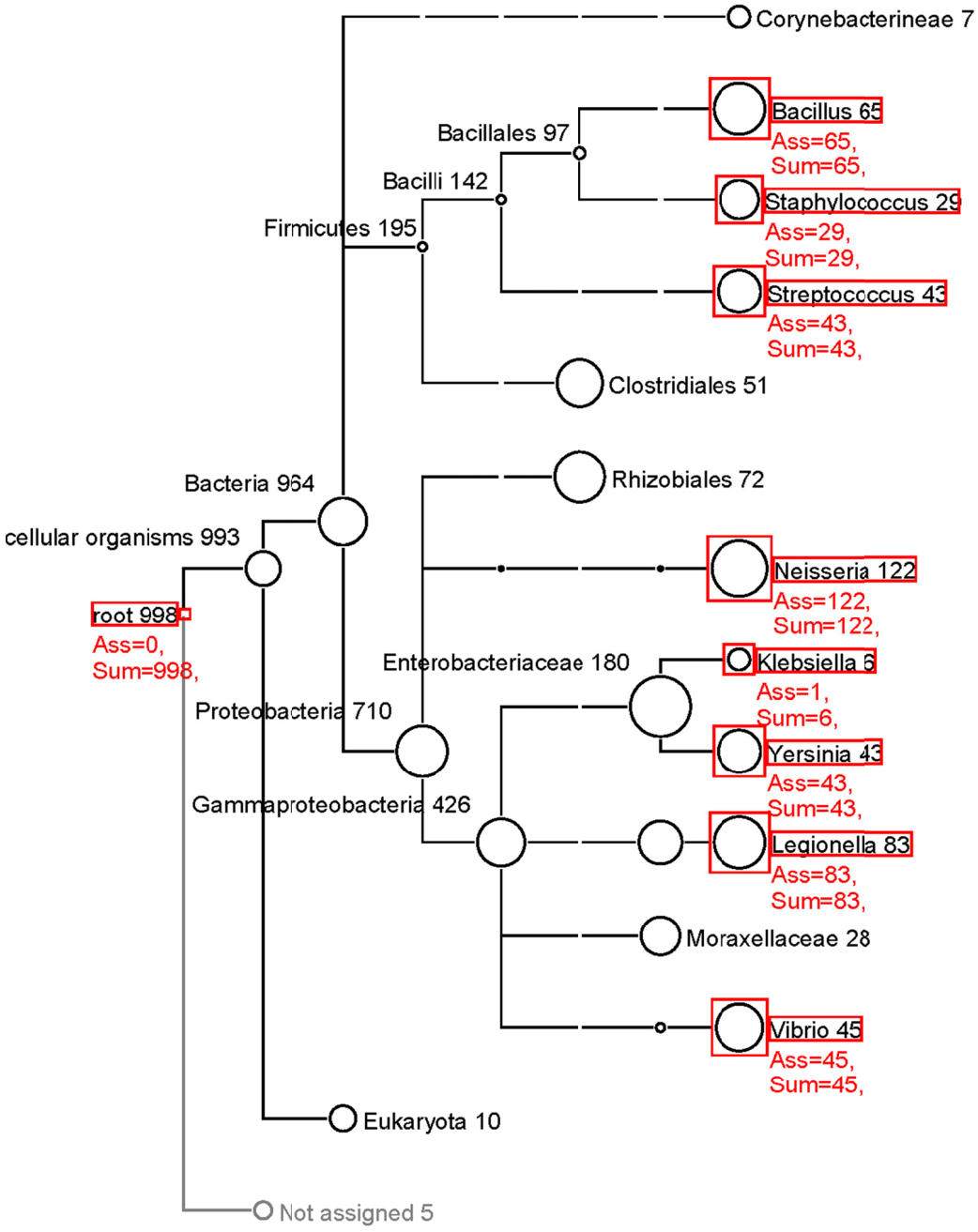


Organisms identified in Sample 24

Species known		Genus identified		
Species (ID)	**%***	**Genus**	**454 %**	**Illumina %**
Klebsiella pneumoniae (8, 10, 11, 28)	**19.2**	**Klebsiella**	**15.52**	**0.44**
Escherichia coli (2)	**4.8**	**Escherichia**	**0.30**	
Bacillus thuringiensis (3)	**4.8**	**Bacillus**	**3.39**	**5.84**
Acinetobacter baumannii (6)	**4.8**	**Acinetobacter**	**7.18**	
Neisseria meningitides (7)	**4.8**	**Neisseria**	**0.46**	**10.76**
Salmonella enterica (9)	**4.8**	**Salmonella**		
Yersinia enterocolitica (12)	**4.8**	**Yersinia**	**3.49**	**3.55**
Yersinia pseudotuberculosis (27)	**4.8**			
Mycobacterium tuberculosis (16)	**4.8**	**Mycobacterium**		
Streptococcus pneumoniae (18)	**4.8**	**Streptococcus**	**5.92**	**4.69**
Vibrio harveyi (19)	**4.8**	**Vibrio**	**3.54**	**3.84**
Rhizobium radiobacter (20)	**4.8**	**Rhizobium**	**0.30**	
Clostridium difficile (21)	**4.8**	**Clostridium**		
Staphylococcus aureus (22, 26)	**4.8**	**Staphylococcus**	**17.34**	**3.20**
Ochrobactrum anthropui (23)	**4.8**	**Ochrobactrum**	**0.66**	
Legionella pneumonphia (24)	**4.8**	**Legionella**	**2.63**	**8.87**
Shigella dysteria (29)	**4.8**	**Shigella**		
		Peptostreptococcus	0.10	
		Raoultella	0.05	
		Cronobacter	0.05	

* Sum of known percentages = 100.8 %

Genus-level taxonomy tree from Illumina sequence analysis of Sample 24



2.5.9 Group 5

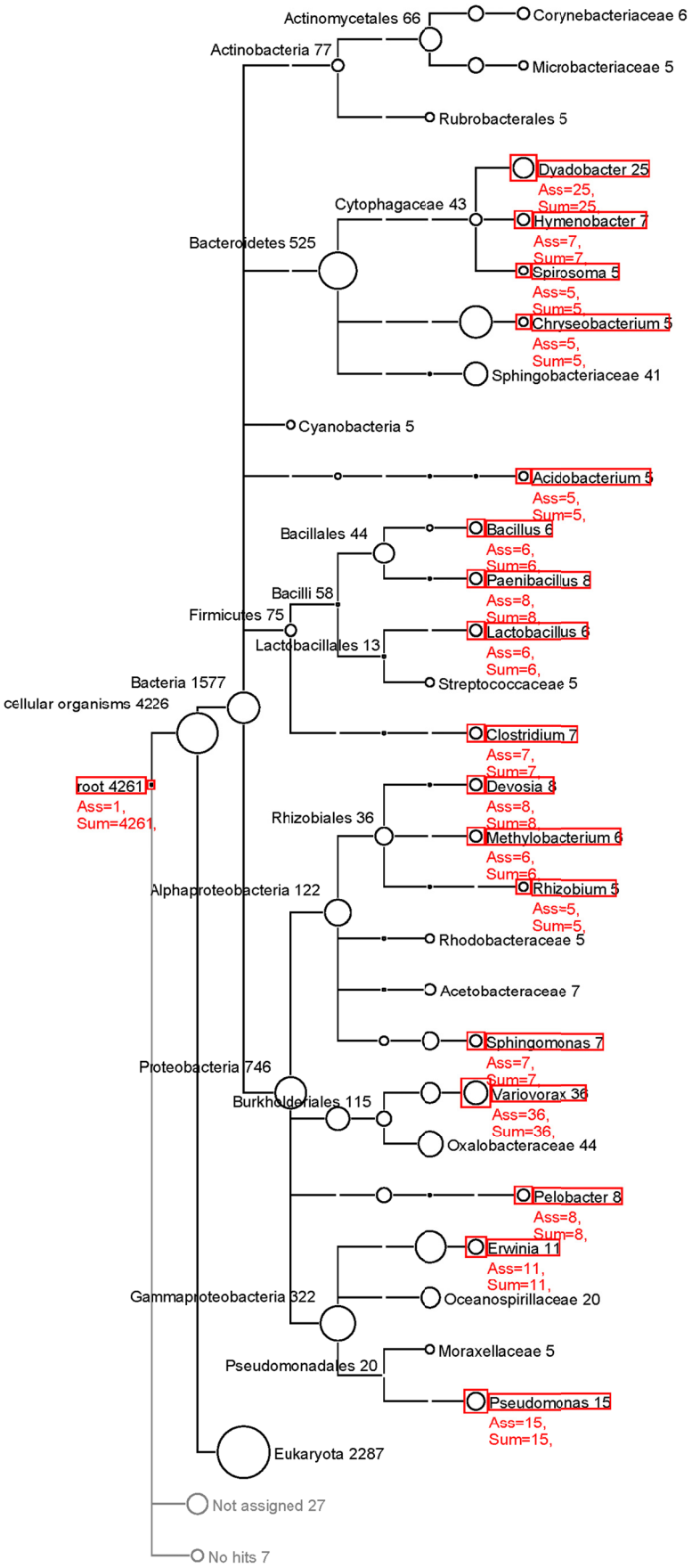
Group 5 consists of Samples 6, 7, 8, and 14. These are environmental PCR samples, where each one is from a different environment: Sample 6 from BK2010, Sample 7 from JBXE2010, Sample 8 from BK2011, and Sample 14 from JBXE2011.

Organisms identified in Samples 6, 7, 8, and 14
(Genera sorted alphabetically. Detection rates greater than 1 % indicated in **boldface**.)

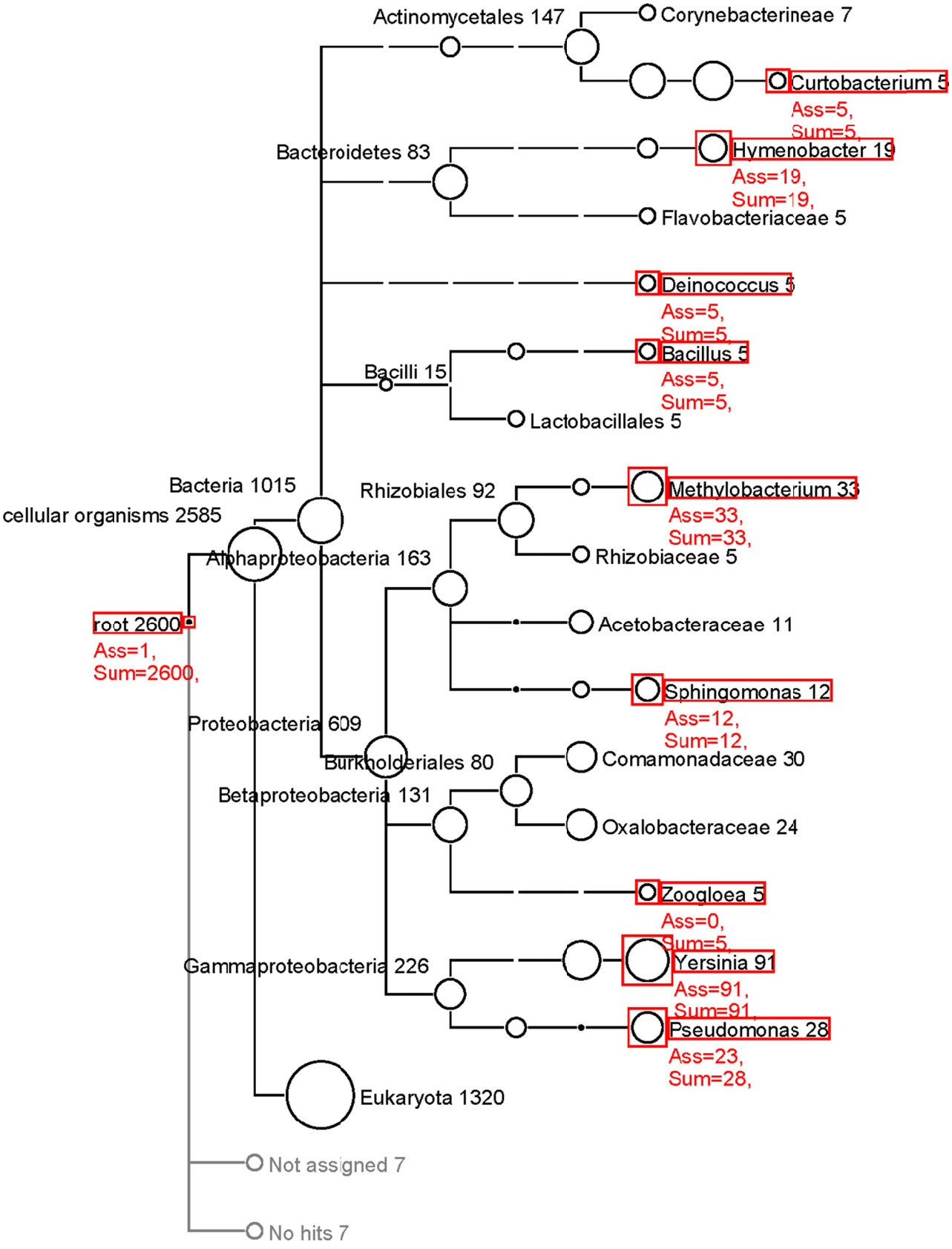
Genus identified	Sample 6		Sample 7		Sample 8		Sample 14	
	454 %	Illumina %	454 %	Illumina %	454 %	Illumina %	454 %	Illumina %
Acidobacterium		0.09				0.31		
Acinetobacter	0.17		0.04		0.97	0.10	**1.68**	**0.20**
Arthrobacter						0.34		0.20
Bacillus		0.13		0.58	**2.1**	**1.92**	**3.47**	**4.48**
Chryseobacterium	**3.26**	0.05			0.23		0	
Clostridium		0.41				0.21		0.56
Columnosphaeria	0.39		0.46		0.45		**1.51**	
Corynebacterium						0.16		
Cryptococcus	0.11		**3.60**				0.29	
Curtobacterium				0.11				
Deinococcus				0.18		0.22		0.79
Devosia		0.06						
Dyadobacter		0.43						
Erwinia	**2.05**	0.15	0.04	0.71	**11.83**	**0.38**		0.50
Flavisolibacter						0.05		
Flavobacterium	**9.02**							
Gemmatimonas						0.35		
Hymenobacter		0.14		0.50				
Klebsiella							**50.93**	
Lactobacillus		0.12				0.37		
Legionella								**1.73**
Massilia	0.55		**1.80**				**2.55**	
Methylobacterium		0.30		0.97		0.17		**1.73**
Mycobacterium						0.23		**1.75**
Neisseria								**1.06**
Nesticus	**1.22**							
Nocardiopsis					**1.95**			
Ochrobactrum								**3.73**
Paenibacillus		0.45				0.17		
Pantoea	**1.27**		0.35		0.23			
Pedobacter	**4.59**		0.19					
Pelobacter		0.12						
Phlebotomus	**1.16**				**1.57**			
Pseudomonas	**14.49**		**1.91**		**5.31**	**1.58**	**4.05**	
Rhizobium		0.11						
Serratia	0.28				**16.92**	**0.23**		

Sphingomonas	0.5	0.25	**1.34**	**0.32**	0.15		0.41	
Spirosoma		0.14						
Sporidiobolus	**1.16**		**1.53**		0.9		0.52	
Sporosarcina					**2.25**			
Staphylococcus							**2.03**	**1.02**
Stenotrophomonas	**1.33**		0.08				0.06	
Streptococcus								**1.50**
Streptomyces						0.13		
Variovax		0.88						0.23
Vibrio								0.95
Yersinia			**62.98**	**2.13**			0.52	0.71
Zoogloea				0.08				

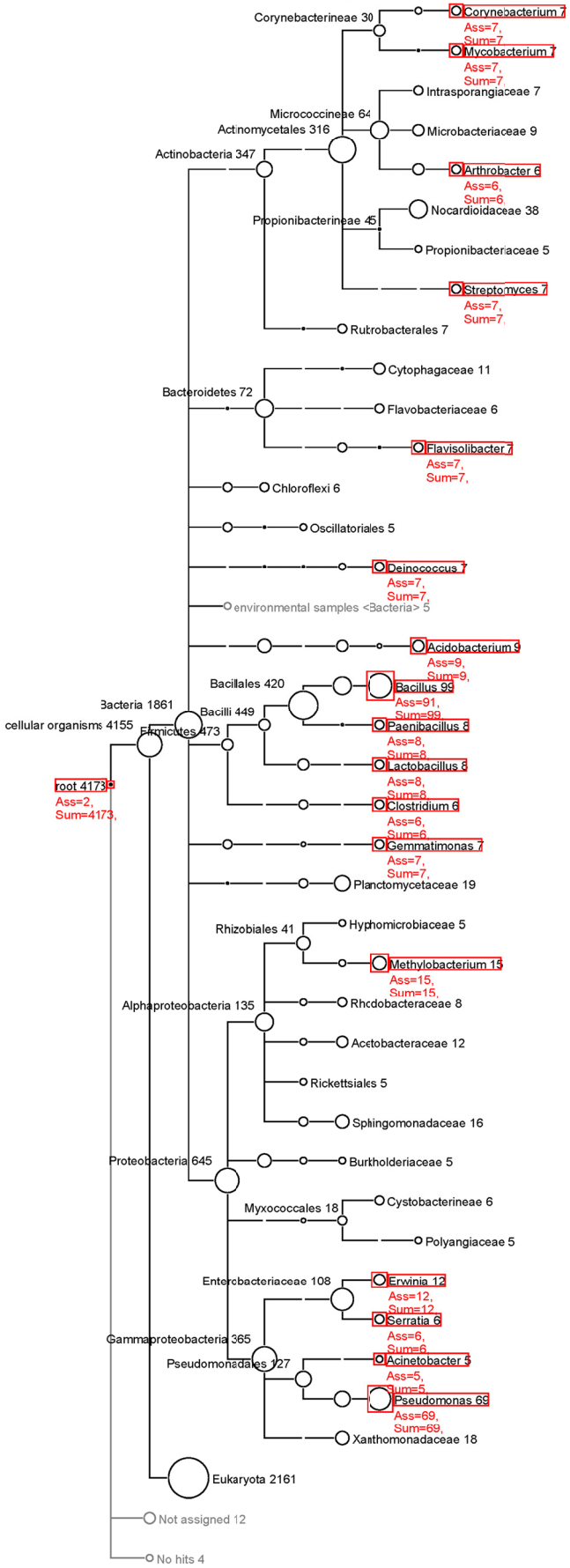
Genus-level taxonomy tree from Illumina sequence analysis of Sample 6



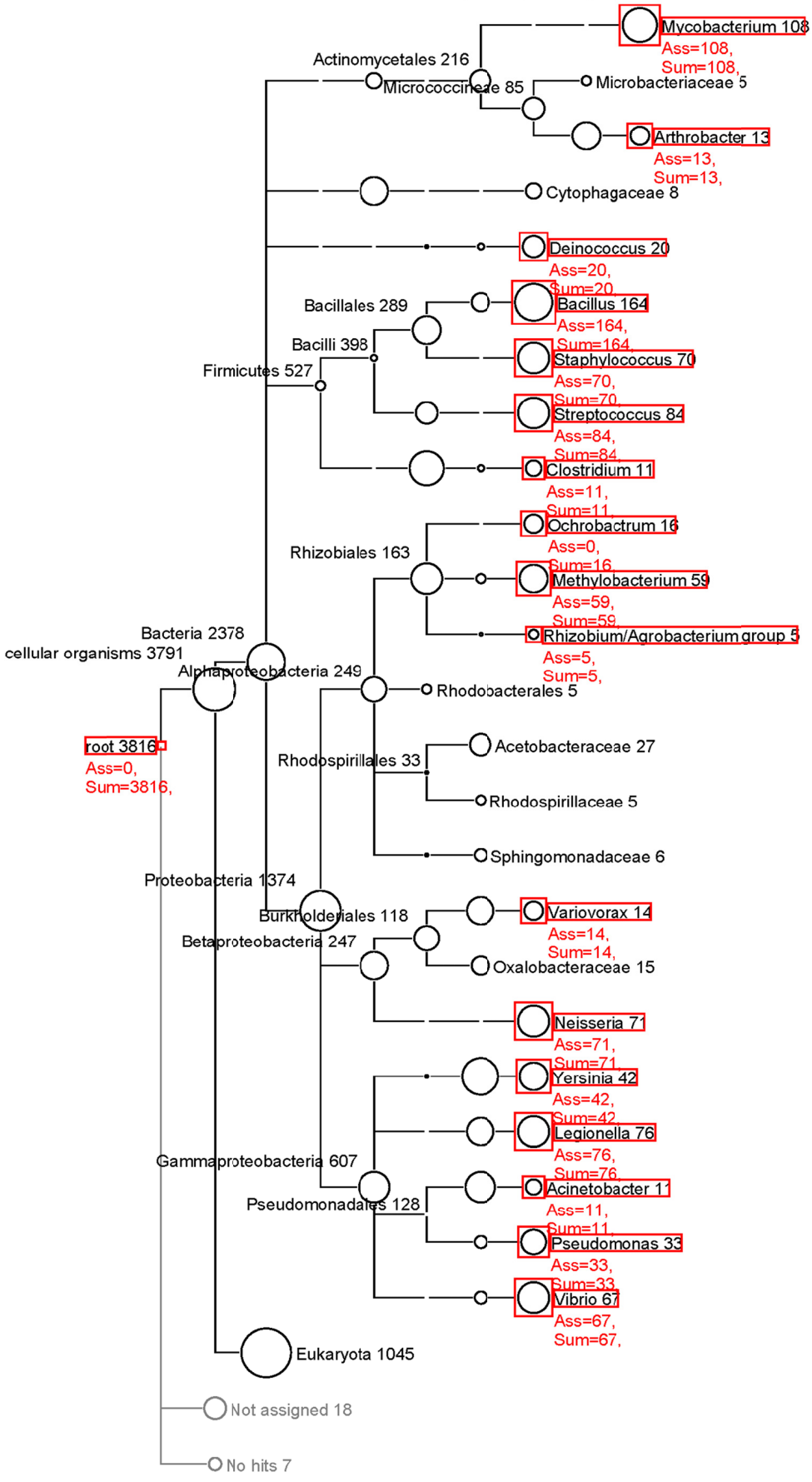
Genus-level taxonomy tree from Illumina sequence analysis of Sample 7



Genus-level taxonomy tree from Illumina sequence analysis of Sample 8



Genus-level taxonomy tree from Illumina sequence analysis of Sample 14



3 Analysis of metagenomic DNA sequences

3.1 Summary

Objective

Whole metagenomic sequencing of microbial DNA samples (long read and short read sequencing) will be investigated to determine if virulence factors or synthetic / genetically modified organisms can be identified using this approach.

Conclusion

We were able to quantify relative abundance of organisms (both at the genus level and at the species level) in the 8 samples using metagenomic DNA sequences. Both the 454 sequences and the Illumina sequences have been analyzed using BLAST and MEGAN. Both approaches succeeded in detecting most known organisms in a sample. For the samples with known relative abundance, we checked the accuracy of the identified relative abundance. In both approaches, there exists a wide range of similarity between the identified and known proportions. For example, at the genus level *Staphylococcus aureus* was identified to comprise 74.49 % (known = 72.7 %) of Sample 1, with 454 sequences. *Bacillus thuringiensis* was identified to comprise 8.68 % (known = 10 %) of the same sample, with Illumina sequences. Identification of other organisms in the same sample does not necessarily produces this level of accuracy.

We were also able to identify a number of markers in the 8 samples. For maker identification, the 454 sequences have been analyzed using BLAST, while the Illumina sequences have been analyzed using Bowtie. Both approaches succeeded in detecting most of the known markers. For example, in Sample 3, 39 (out of 48) markers have been detected. In the same sample, the top 3 markers (in % of reads) from both approaches are the same.

3.2 Grouping of samples

The 8 metagenomic samples were divided into three groups based on whether the organisms (and their proportions) in the samples are known:

- Known: Samples in this group each contain a known set of organisms with known proportions.
- Partially known: Samples in this group each contain a known organism with a known proportion, along with unknown organisms from an environment
- Unknown: Samples in this group are entirely from an environment and thus contain only unknown organisms.

The following table shows the breakdown of the samples according to this grouping.

Group	Sample ID	Composition DNA ID / environment (%)
Known	1	22 (72.7), 29 (11.3), 3 (10), 5 (6)
	2	29 (47), 21 (31), 3 (13), 5 (9)
	3	29 (51), 14 (44), 3 (3.5), 5 (1.8)
	4	27 (55), 26 (23), 19 (22)
Partially known	7	27 (76), environment JBYE 2010 (24)
	14	27 (70), environment JBYE 2010 (30)
Unknown	6	Environment BK 2010 (100)
	8	Environment BK 2011 (100)

The organisms known to be contained in each sample are as follows:

- Sample 1:
  - *Bacillus thuringiensis*, *Acinetobacter baumannii*, *Straphylococcus aureus*, *Shigella dysteria* (4 organisms)
- Sample 2:
  - *Bacillus thuringiensis*, *Acinetobacter baumannii*, *Clostridium difficile*, *Shigella dysteria* (4 organisms)
- Sample 3:
  - *Bacillus thuringiensis*, *Acinetobacter baumannii*, *Klebsiella pneumoniae*, *Shigella dysteria* (4 organisms)
- Sample 4:
  - *Vibrio harveyi*, *Straphylococcus aureus*, *Yersinia pseudotuberculosis* (3 organisms)
- Sample 7: *Yersinia pseudotuberculosis* plus unknown organisms from JBYE 2010
- Sample 14: *Yersinia pseudotuberculosis* plus unknown organisms from JBYE 2011
- Sample 6: Unknown organisms from BK 2010
- Sample 8: Unknown organisms from BK 2011

3.3 Processing of 454 sequences

Based on a FastQC quality pre-check, the quality of raw reads from each sample was checked as follows:

- Trim from the end until the quality score of the last base is at least 20
- Filter if the read length (after trimming) is less than 100, there is an ambiguous (N) base, or the average quality score over all bases is less than 25

The good reads were then subsampled to 10%. This was done because there were simply too many reads to run the analysis in a reasonable amount of time. The subsampling uses a random number generator. This guarantees that the subset of the raw reads selected by the random selection process represents the full set of reads without bias. The following table shows the outcome of the quality control and subsampling processes.

Sample ID	Raw reads	Good reads	Subsampled reads
1	351,131	277,379 (79 %)	27,738
2	346,576	264,646 (76 %)	26,465
3	299,023	242,042 (81 %)	24,204
4	443,015	357,501 (81 %)	35,750
6	350,204	272,851 (78 %)	27,285
7	329,352	279,921 (85 %)	27,992
8	305,143	235,169 (77 %)	23,517
14	402,280	333,063 (83 %)	33,306

For the analysis of **relative abundance of organisms**, BLAST (blastn) was run on the subsampled reads for each sample against the NCBI "nt" database. We then used MEGAN to parse the BLAST output and run the taxonomy analysis. The proportion of an identified taxon, which is to be compared with the known proportion, was calculated by the number of reads assigned to the taxon.

For the identification of **virulence factors or synthetic / genetically modified markers**, BLAST (blastn) was run on the whole set of good reads (i.e. without subsampling) against the 48 unique target marker sequences. The top hits were collected to profile the identification results.

3.4 Processing of Illumina sequences

The overall quality of the Illumina metagenomic samples were first checked using the FastQC program. The FastQC quality pre-check showed that the quality of the R2 reads were generally poorer than that of R1 reads. Therefore, we used a more relaxed set of quality control parameters for R2 sequences. This prevented many perfectly good R1 reads from being abandoned after quality control, as a result of failing to pair with matching R2 reads which could have been filtered out. The final parameters for R1 reads were set as follows:

- Trim from the end until the quality score of the last base is at least 20
- Filter if the read length (after trimming) is less than 50 for R1, there is an ambiguous (N) base, or the average quality score over all bases is less than 27

For R2 reads they were set relatively leniently as follows:

- Trim from the end until the quality score of the last base is at least 15
- Filter if the read length (after trimming) is less than 15 for R1, there is an ambiguous (N) base, or the average quality score over all bases is less than 20

For the assembly of the short reads, we first used the Velvet assembler. After experimenting with a few different hash lengths for the assembly, we chose 67 as the *k*-mer length to use. We then used MetaVelvet on the Velvet output files, which is an extension of Velvet recently developed to postprocess Velvet output for handling of metagenomic reads. The results of the quality control using the above parameters and the assembly using Velvet followed by MetaVelvet are summarized in the following table.

Sample ID	Quality control			Assembly using Velvet and MetaVelvet (Hash length 67, insert length 450)				
	R1/R2	Raw reads	Good reads	Number of all contigs	N50	Max contig length	Total assembly length	Number of high-coverage contigs used for analysis
1	R1	21,310,900	21,027,236 (98.7 %)	5,860	18,039	183,610	16,757,248	3,901
	R2	21,310,900	20,562,559 (96.5 %)					
	Paired		40,692,594					
2	R1	30,241,450	29,868,600 (98.8 %)	4,168	60,513	354,970	18,716,548	2,003
	R2	30,241,450	29,128,594 (96.3 %)					
	Paired		57,706,286					
3	R1	22,917,844	22,444,166 (98.0 %)	4.028	23,089	166,563	9,896,944	2,042
	R2	22,917,844	21,677,096 (94.6 %)					
	Paired		42,782,174					
4	R1	20,346,292	20,070,524 (98.6 %)	3,441	32,956	150,795	12,935,102	1,635
	R2	20,346,292	19,417,947 (95.4 %)					
	Paired		38,490804					
6	R1	25,649,024	24,835,303 (96.8 %)	83,437	241	2,895	10,569,101	37,235
	R2	25,649,024	24,542,402 (95.7 %)					
	Paired		47,771,906					
7	R1	31,486,879	30,868,442 (98.0 %)	884	36,986	130,862	4,717,635	484
	R2	31,486,879	30,184,100 (95.9 %)					
	Paired		59,468,744					
8	R1	25,956,836	25,221,154 (97.2 %)	13,356	36,367	201,674	6,543,850	3,275
	R2	25,956,836	25,071,348 (96.6 %)					
	Paired		48,946,556					
14	R1	31,559,547	30,650,195 (97.1 %)	976	81,915	201,674	5,670,329	582
	R2	31,559,547	30,258,630 (95.9 %)					
	Paired		59,175,628					

For each sample, a set of high-coverage contigs was used for organism identification. The number of such contigs for each sample is shown in the last column of the above table.

For the analysis of **relative abundance of organisms**, BLAST (blastn) was run on the selected contigs for each sample against the NCBI "nt" database. Then we used MEGAN to parse the BLAST output and run the taxonomy analysis. The proportions of the identified taxa, which are to be compared with the known proportions, were calculated not by the number of contigs assigned to the taxon, but by the sum of the median coverages of all the contigs assigned to the taxon. This way, the number of reads assembled into a contig is accounted for when we estimate the abundance of the organism to which the contig has been assigned.

For the identification of **virulence factors or synthetic / genetically modified markers**, Bowtie was run on the same set of quality-controlled R1 reads against the 48 unique target sequences. The R2 reads were deemed too short to be used for mapping. The best hits were collected to profile the identification results.

3.5 Identification of organisms and estimation of their proportions

For each metagenomic sample, we have performed two separate analyses based on 454 sequences and Illumina sequences, respectively, to identify organisms present in the sample and to estimate their relative abundance. The analysis results for each sample are presented side by side per organism, to facilitate the comparison of the differences between the 454-based and Illumina-based approaches.

For each sample, a table is given to show identified organisms from each analysis and their estimated proportions. In these tables, **red boldface** letters and percentages in the "Species known" column indicate the known organisms and their proportions. The **boldface** letters and percentages in the "Genus identified" and "Species identified" columns indicate the same organisms identified by the two (454 and Illumina) analyses, at the genus level and at the species level, respectively. If a table cell in these two identification columns is empty, it indicates that the particular genus or species was not identified by the analysis. After this table, two taxonomy trees expanded down to the genus level are shown, one from the Illumina sequence analysis and the other from the 454 sequence analysis.

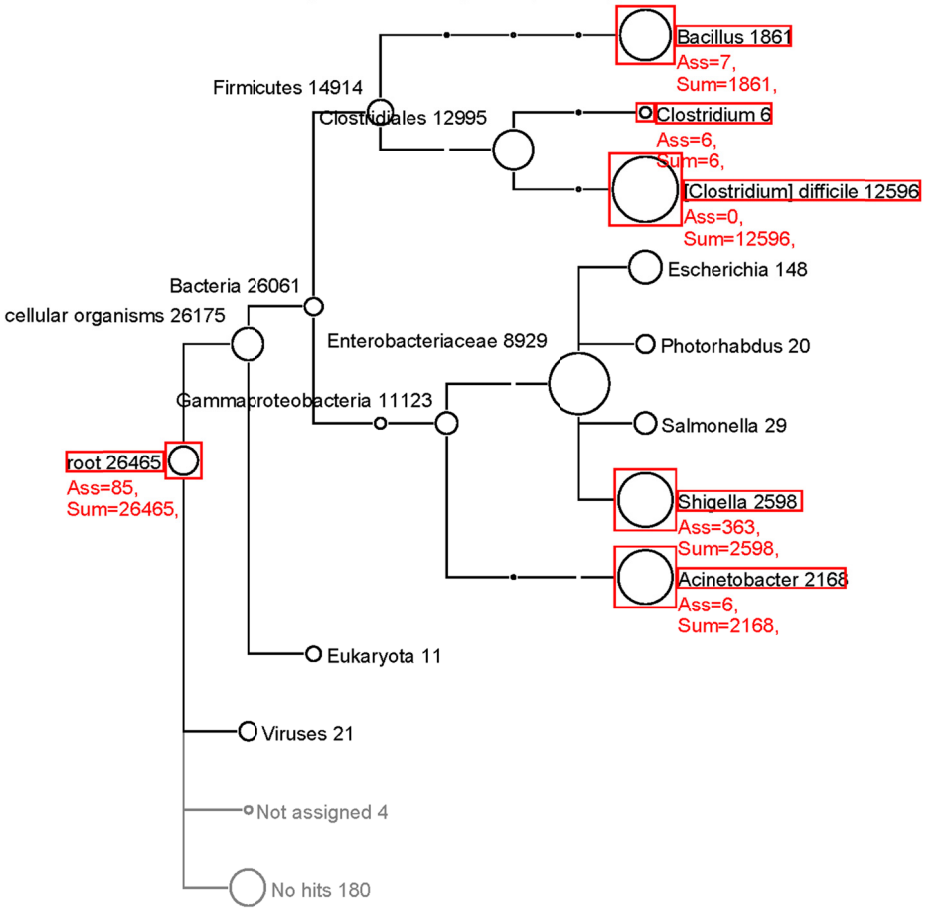
3.5.1 "Known" group

This group consists of Samples 1, 2, 3, and 4.

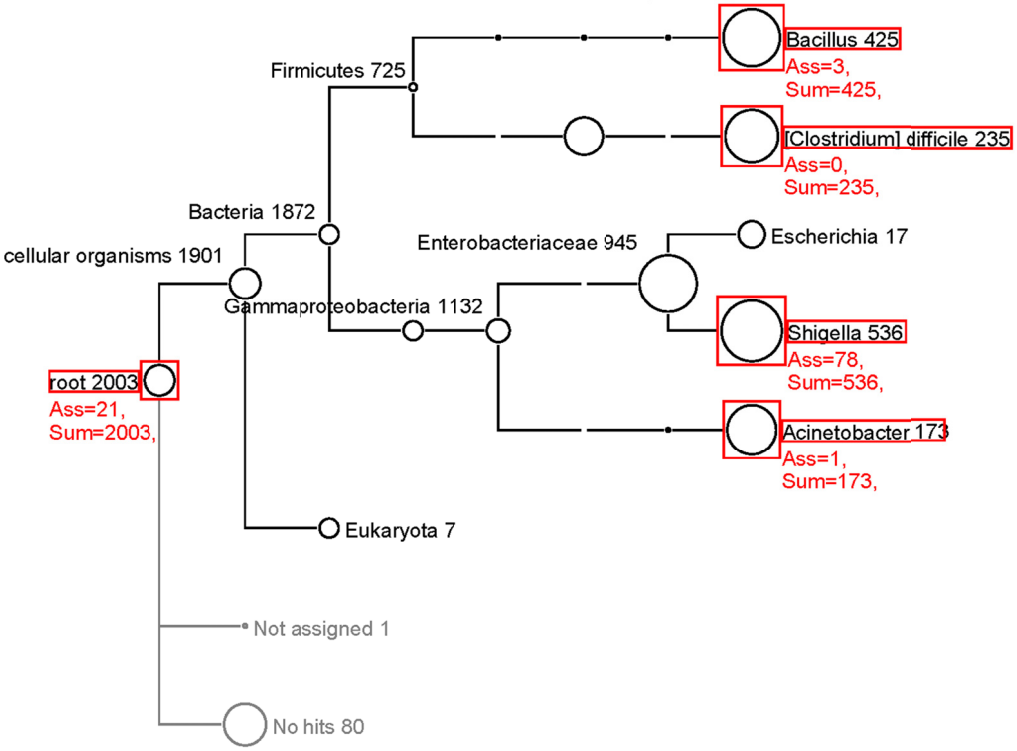
Organisms identified in Sample 1

Species known		Genus identified			Species Identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**	**Species**	**454 %**	**Illumina %**
Staphylococcus aureus (22)	**72.7**	**Staphylococcus**	**74.49**	**49.55**	**Staphylococcus aureus**	**72.53**	**36.55**
					Staphylococcus epidermidis	0.03	
Shigella dysteria (29)	**11.3**	**Shigella**	**3.56**	**8.61**	**Shigella dysteria**	**2.99**	**5.98**
					Shigella sonnei	0.04	
Bacillus thuringiensis (3)	**10.0**	**Bacillus**	**4.87**	**8.68**	**Bacillus thuringiensis**	**0.29**	**1.24**
					Bacillus cereus	3.95	4.13
					Bacillus anthracis		0.12
					Bacillus weihenstephanensis	0.06	0.22
Acinetobacter baumannii (5)	**6.0**	**Acinetobacter**	**5.79**	**11.34**	**Acinetobacter baumannii**	**5.74**	**11.33**
		Escherichia	0.22	0.23	Escherichia coli	0.22	0.23
		Salmonella	0.04		Salmonella enterica	0.04	

Genus-level taxonomy tree from 454 sequence analysis of Sample 1



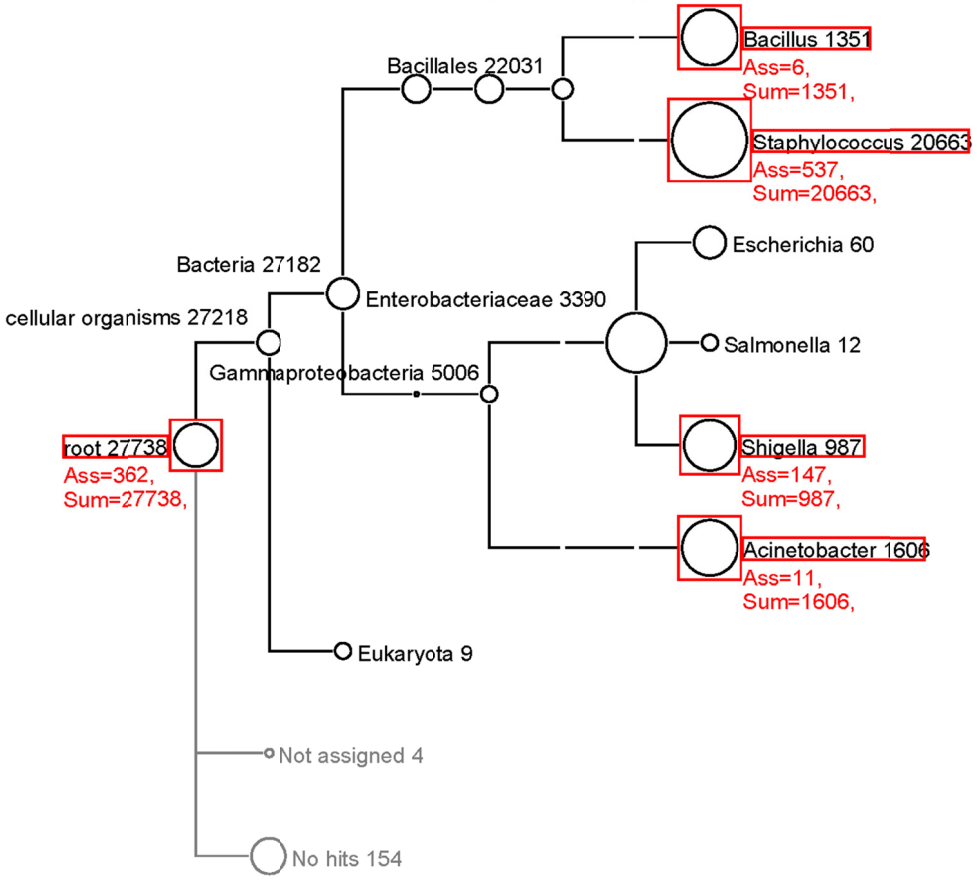
Genus-level taxonomy tree from Illumina sequence analysis of Sample 1



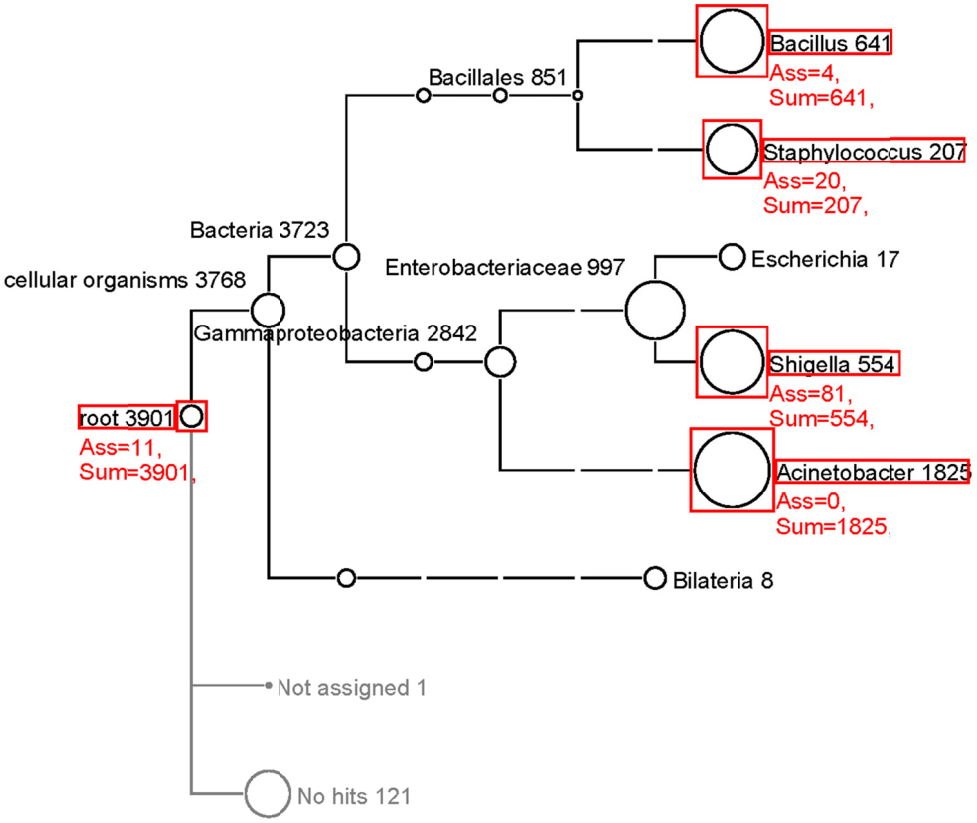
Organisms identified in Sample 2

Species known		Genus identified			Species Identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**	**Species**	**454 %**	**Illumina %**
Shigella dysteria (29)	**47.0**	**Shigella**	**9.82**	**27.43**	**Shigella dysteria**	**8.38**	**20.42**
					Shigella sonnei	0.07	
Clostridium difficile (21)	**31.0**	**Clostridium**	**47.62**	**17.94**	**Clostridium difficile**	**47.59**	**17.94**
Bacillus thuringiensis (3)	**13.0**	**Bacillus**	**7.03**	**7.52**	**Bacillus thuringiensis**	**0.46**	**1.66**
					Bacillus cereus	5.85	2.84
					Bacillus weihenstephanensis	0.10	0.17
					Bacillus anthracis	0.02	0.12
Acinetobacter baumannii (5)	**9.0**	**Acinetobacter**	**8.19**	**2.55**	**Acinetobacter baumannii**		**2.49**
		Escherichia	0.56	0.63	Escherichia coli	0.56	0.63
		Salmonella	0.11		Salmonella enterica	0.11	
		Photorhabdus	0.08		Photorhabdus luminescens	0.08	

Genus-level taxonomy tree from 454 sequence analysis of Sample 2



Genus-level taxonomy tree from Illumina sequence analysis of Sample 2

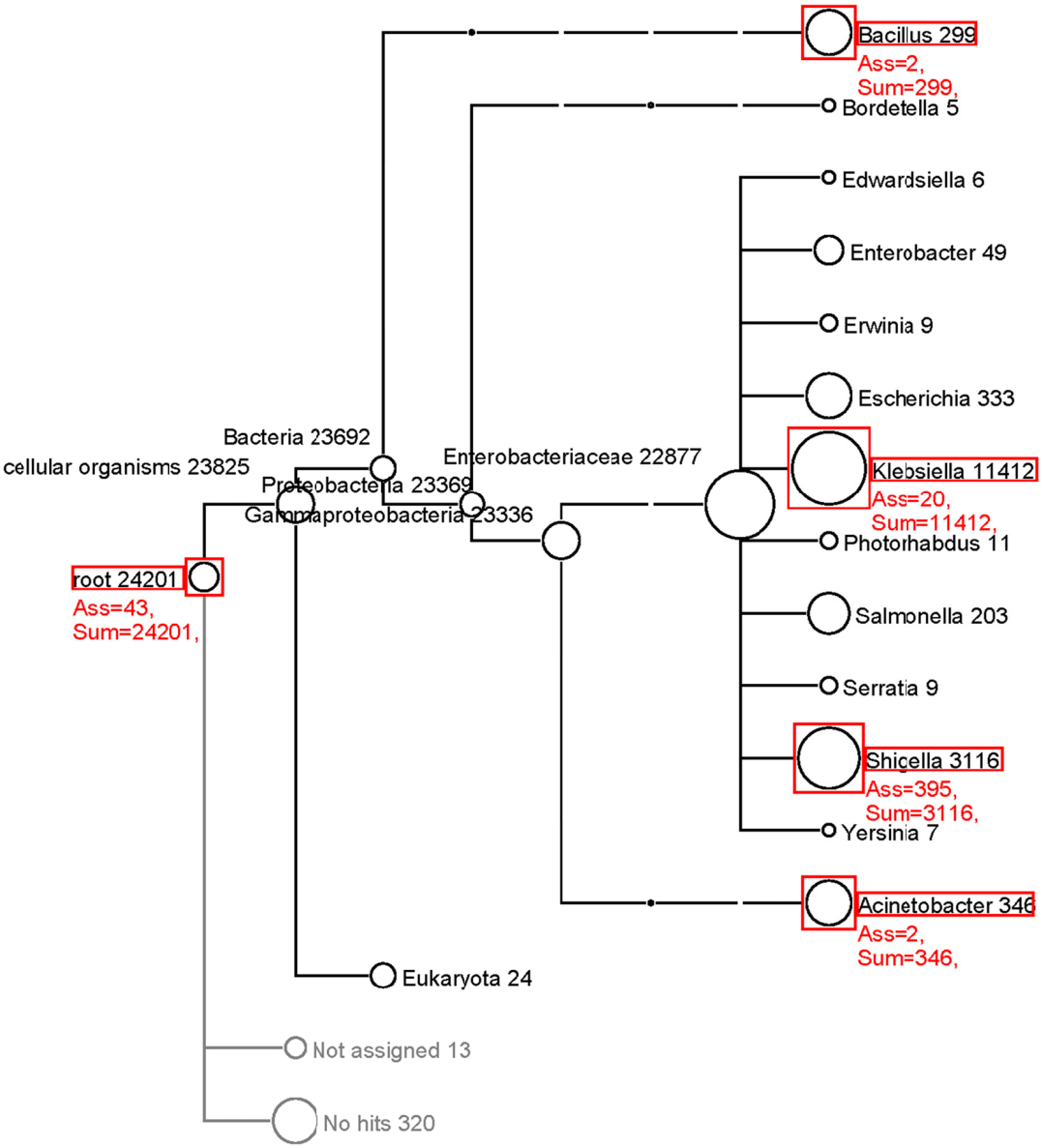


Organisms identified in Sample 3

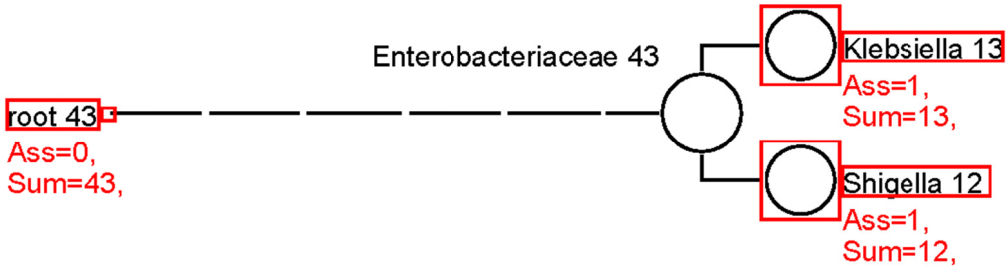
Species known		Genus identified			Species Identified		
Species (ID)	**%***	**Genus**	**454 %**	**Illumina %**	**Species**	**454 %**	**Illumina %**
Shigella dysteria (29)	**51.0**	**Shigella**	**12.88**	**8.47**	**Shigella dysteria**	**11.04**	**7.27**
					Shigella sonnei	0.14	
					Shigella boydii	0.06	
Klebsiella pneumoniae (14)	**44.0**	**Klebsiella**	**47.16**	**30.02**	**Klebsiella pneumoniae**	**47.07**	**7.60**
Bacillus thuringiensis (3)	**3.5**	**Bacillus**	**1.24**		**Bacillus thuringiensis**	**0.06**	
					Bacillus cereus	1.04	
Acinetobacter baumannii (5)	**1.8**	**Acinetobacter**	**1.43**		**Acinetobacter baumannii**	**1.42**	
		Escherichia	1.38		Escherichia coli	1.37	
		Salmonella	0.84		Salmonella enterica	0.82	
		Enterobacter	0.20				
		Photorhabdus	0.05		Photorhabdus luminescens	0.05	
		Serratia	0.04		Serratia proteamaculans	0.04	
		Erwinia	0.04		Erwinia tasmaniensis	0.02	
		Yersinia	0.03				
		Edwardsiella	0.02		Edwardsiella ictaluri	0.02	
		Bordetella	0.02				

* Sum of known percentages = 100.3 %

Genus-level taxonomy tree from 454 sequence analysis of Sample 3



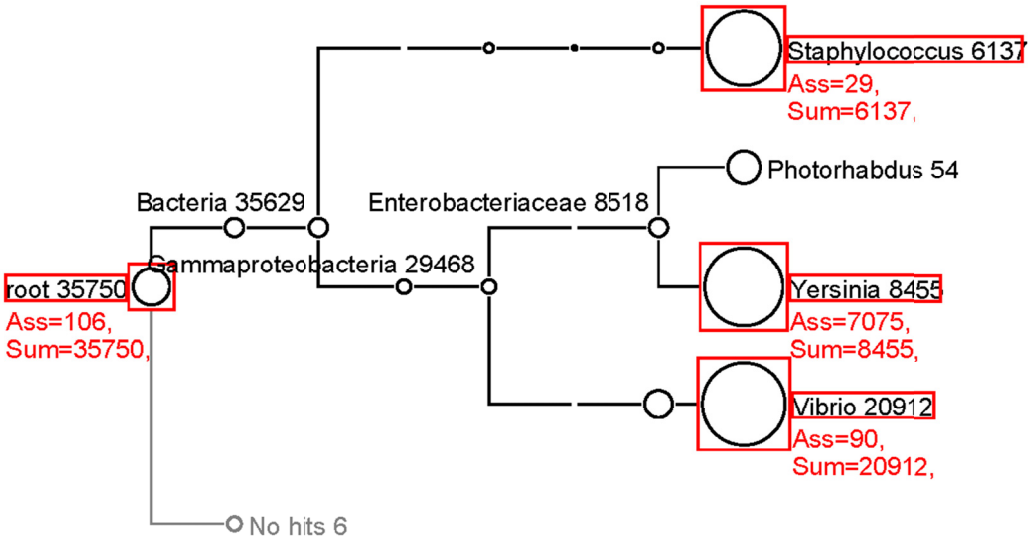
Genus-level taxonomy tree from Illumina sequence analysis of Sample 3



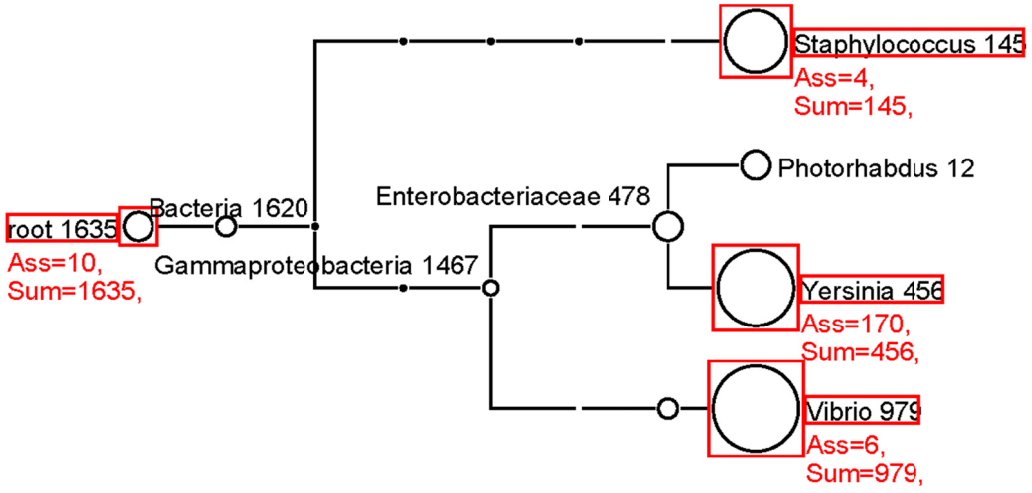
Organisms identified in Sample 4

Species known		Genus identified			Species Identified		
Species (ID)	**%**	**Genus**	**454 %**	**Illumina %**	**Species**	**454 %**	**Illumina %**
Yersinia pseudotuberculosis (27)	**55.0**	**Yersinia**	**23.65**	**4.04**	**Yersinia pseudotuberculosis**	**3.79**	**1.82**
					Yersinia pestis	0.07	
Staphylococcus aureus (26)	**23.0**	**Staphylococcus**	**17.17**	**3.00**	**Staphylococcus aureus**	**17.09**	**2.82**
Vibrio harveyi (19)	**22.0**	**Vibrio**	**58.50**	**88.84**	**Vibrio harveyi**	**57.85**	**85.75**
		Photorhabdus	0.15	0.29	Photorhabdus luminescens	0.15	0.29

Genus-level taxonomy tree from 454 sequence analysis of Sample 4



Genus-level taxonomy tree from Illumina sequence analysis of Sample 4



3.5.2 "Partially known" group

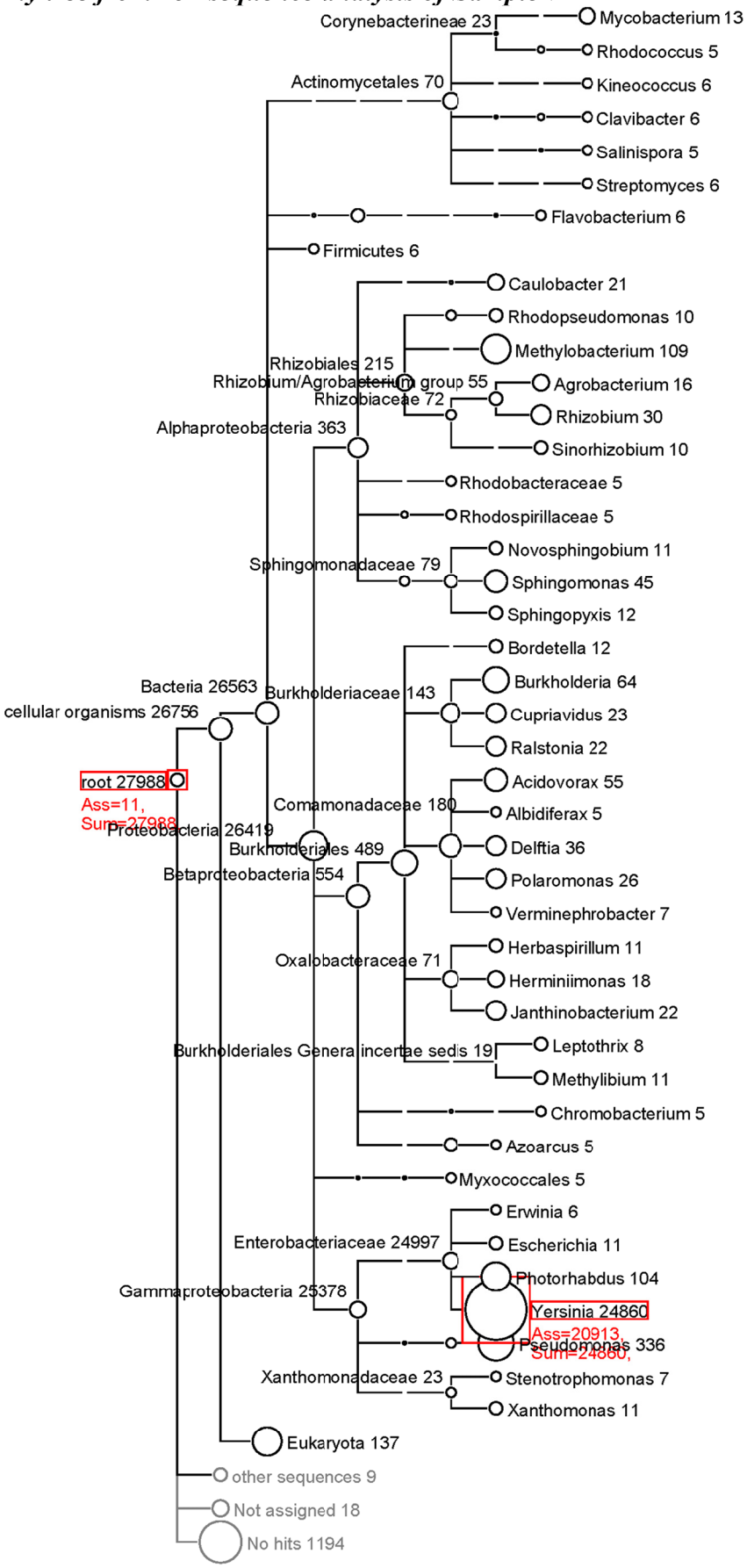
This group consists of Samples 7 and 14. Both samples are known to contain one organism ***Yersinia pseudotuberculosis*** in different proportions: **76 %** in Sample 7 and **70 %** in Sample 14.

Organisms identified in Samples 7 and 14
(Genera sorted in descending order of abundance for the known organism and for those with a detection rate greater than 1 %, indicated in **boldface**. Other genera listed alphabetically.)

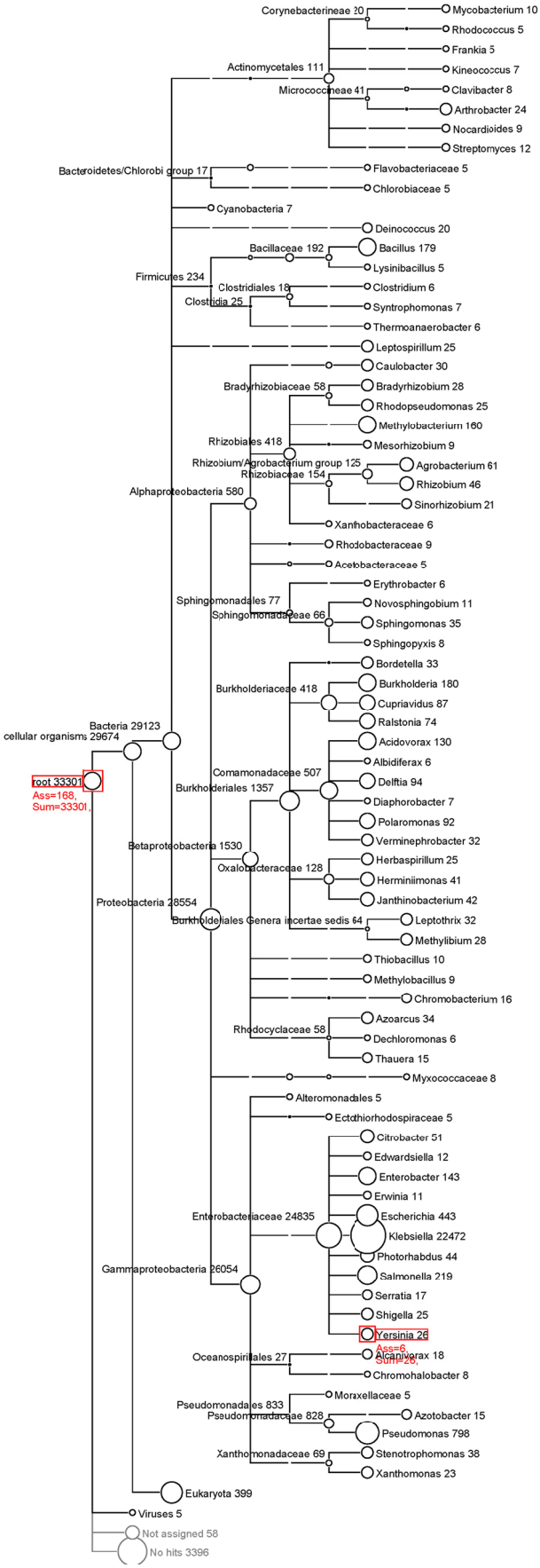
Genus identified	Sample 7		Sample 14	
	454 %	Illumina %	454 %	Illumina %
Yersinia	**88.82**	**96.40**	**0.08**	
Yersinia pseudotuberculosis	*13.77*	*44.32*		
Klebsiella			**67.48**	**45.45**
Pseudomonas	**1.20**		**2.40**	
Salmonella			0.66	**1.61**
Escherichia	0.04		**1.33**	0.95
Acidovorax	0.20		0.39	
Actinomycetales			0.33	
Agrobacterium	0.06		0.18	
Albidiferax	0.02		0.02	
Alcanivorax			0.05	
Arthrobacter			0.07	
Azoarcus	0.02		0.10	
Azotobacter			0.05	
Bacillus			0.54	
Bordetella	0.04		0.10	
Bradyrhizobium			0.08	
Burkholderia	0.23		0.54	
Caulobacter	0.08		0.09	
Chromobacterium	0.02		0.05	
Chromohalobacter			0.02	
Citrobacter			0.15	
Clavibacter	0.02		0.02	
Clostridium			0.02	
Cupriavidus	0.08		0.26	
Dechloromonas			0.02	
Deinococcus			0.06	
Delftia	0.13		0.28	
Diaphorobacter			0.02	
Edwardsiella			0.04	
Enterobacter			0.43	
Erwinia	0.02		0.03	
Erythrobacter			0.02	
Flavobacterium	0.02			
Frankia			0.02	
Herbaspirillum	0.04		0.08	

Herminiimonas	0.06		0.12	
Janthinobacterium	0.08		0.13	
Kineococcus	0.02		0.02	
Leptospirillum			0.08	
Leptothrix	0.03		0.10	
Lysinibacillus			0.02	
Mesorhizobium			0.03	
Methylibium	0.04		0.08	
Methylobacillus			0.03	
Methylobacterium	0.39		0.48	
Mycobacterium	0.05		0.03	
Nocardioides			0.03	
Novosphingobium	0.04		0.03	
Photorhabdus	0.37		0.13	
Polaromonas	0.09		0.28	
Ralstonia	0.08		0.22	
Rhizobium	0.11		0.14	
Rhodococcus	0.02		0.02	
Rhodopseudomonas	0.04		0.08	
Salinispora	0.02			
Serratia			0.05	
Shigella			0.08	
Sinorhizobium	0.04		0.06	
Sphingomonas	0.16		0.11	
Sphingopyxis	0.04		0.02	
Stenotrophomonas	0.03		0.11	
Streptomyces	0.02		0.04	
Syntrophomonas			0.02	
Thauera			0.05	
Thermoanaerobacter			0.02	
Thiobacillus			0.03	
Verminephrobacter	0.03		0.10	
Xanthomonas	0.04		0.07	

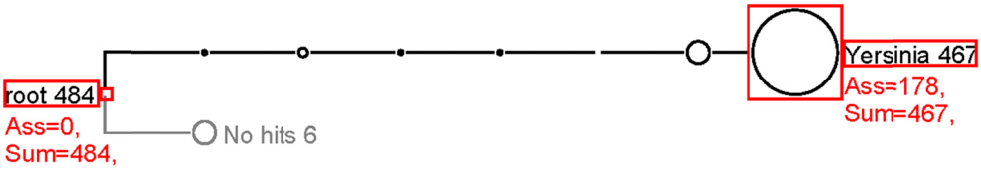
Genus-level taxonomy tree from 454 sequence analysis of Sample 7



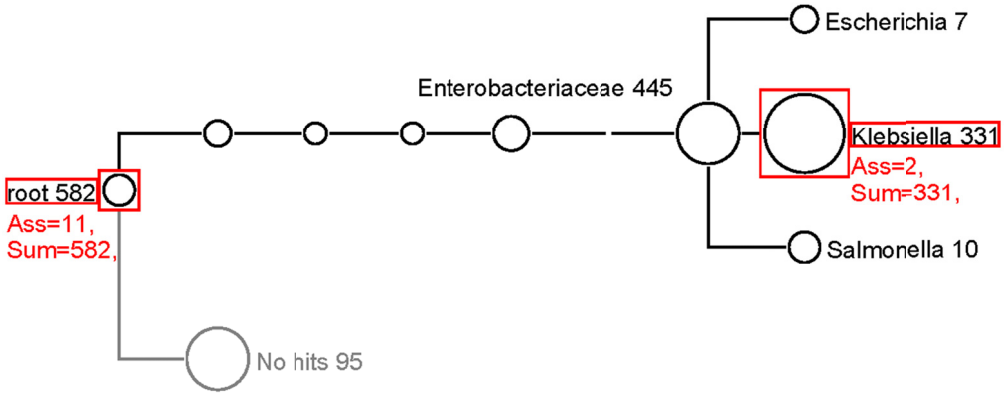
Genus-level taxonomy tree from 454 sequence analysis of Sample 14



Genus-level taxonomy tree from Illumina sequence analysis of Sample 7



Genus-level taxonomy tree from Illumina sequence analysis of Sample 14



3.5.3 "Unknown" group

This group consists of Samples 6 and 8.

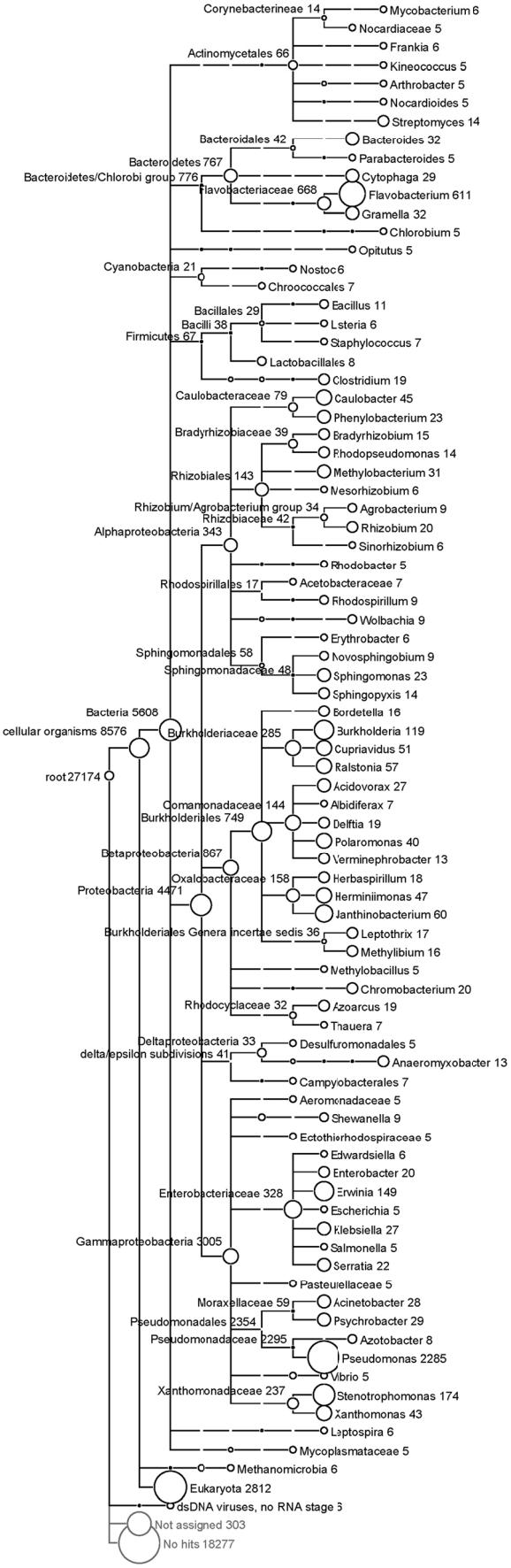
Organisms identified in Samples 6 and 8

(Genera sorted in descending order of abundance for those with a detection rate greater than 1 %, indicated in **boldface**. Other genera listed alphabetically.)

Genus identified	Sample 6		Sample 8	
	454 %	Illumina %	454 %	Illumina %
Pseudomonas	**8.41**	**21.58**	0.73	
Klebsiella	0.10	0.35	0.05	**8.84**
Erwinia	0.55	**3.81**	0.57	
Flavobacterium	**2.25**	0.12		
Serratia	0.08	0.43	**2.01**	0.09
Acidithiobacillus		0.01		
Acidovorax	0.10	0.02		
Acinetobacter	0.10		0.20	
Aeromonas		0.01		
Agrobacterium	0.03	0.02		
Albidiferax	0.03			
Anaeromyxobacter	0.05		0.02	
Arthrobacter	0.02			
Azoarcus	0.07			
Azotobacter	0.03	0.03		
Bacillus	0.04		0.11	
Bacteroides	0.12			
Bordetella	0.06	0.01		
Bradyrhizobium	0.06		0.02	
Burkholderia	0.44	0.13	0.07	
Caulobacter	0.17			
Chlorobium	0.02			
Chromobacterium	0.07	0.01		
Citrobacter		0.05		
Clostridium	0.07	0.01	0.04	
Cupriavidus	0.19	0.03		
Cytophaga	0.11			
Delftia	0.07	0.02		
Edwardsiella	0.02	0.11	0.02	
Enterobacter	0.07		0.05	
Erythrobacter	0.02			
Escherichia	0.02	0.10	0.03	0.25
Frankia	0.02			
Gemmatimonas			0.02	
Gramella	0.12			
Herbaspirillum	0.07			
Herminiimonas	0.17	0.02		

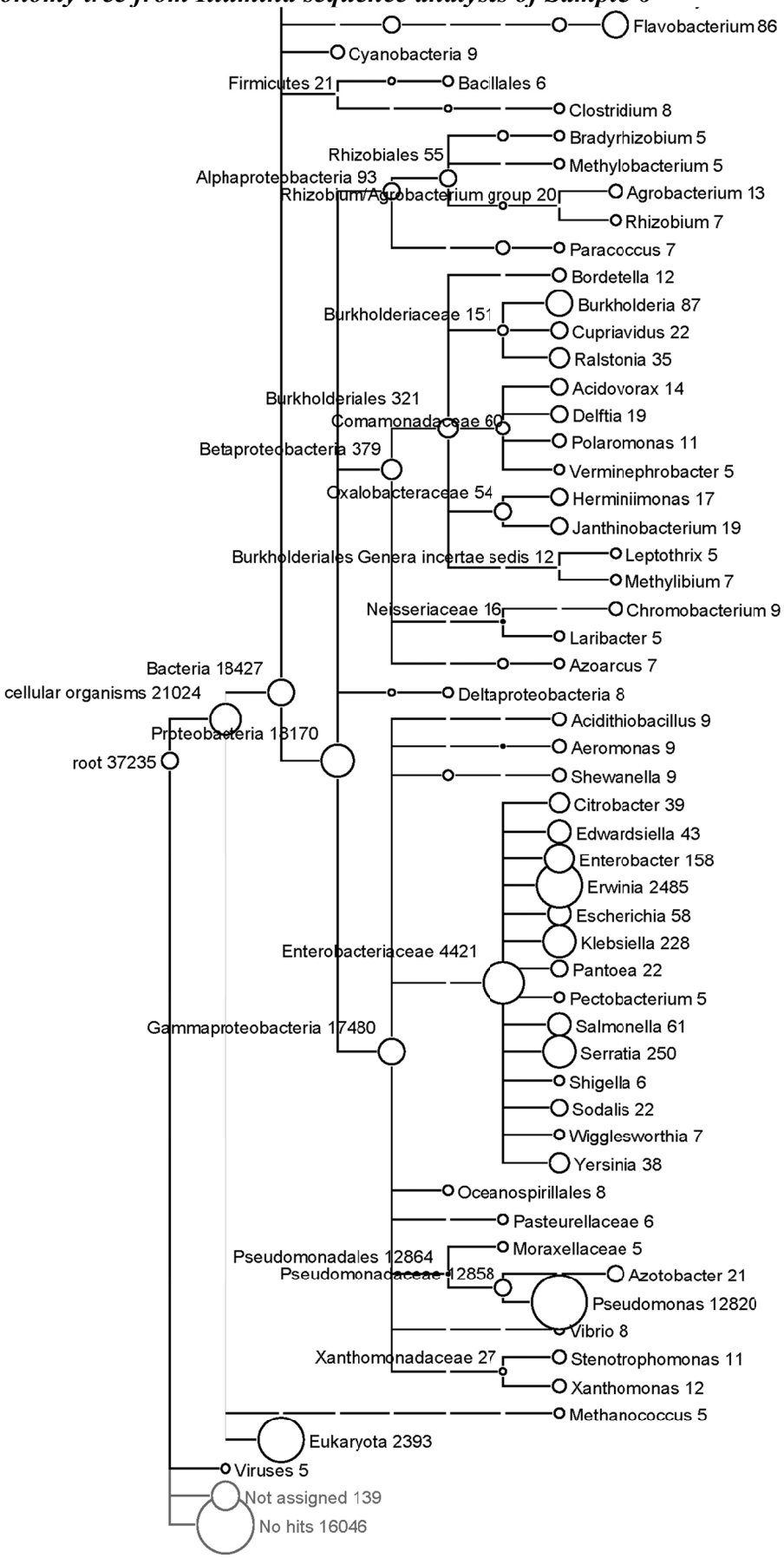
Janthinobacterium	0.22	0.02		
Kineococcus	0.02			
Leptospira	0.02			
Leptothrix	0.06			
Listeria	0.02			
Lysinibacillus			0.04	
Mesorhizobium	0.02			
Methanococcus		0.01	0.03	
Methylibium	0.06	0.01		
Methylobacillus	0.02			
Methylobacterium	0.11		0.03	
Mycobacterium	0.02	0.01	0.03	
Nocardioides	0.02		0.04	
Nostoc	0.02			
Novosphingobium	0.03			
Oceanobacillus			0.03	
Opitutus	0.02			
Pantoea		0.06		
Parabacteroides	0.02			
Paracoccus		0.01		
Phenylobacterium	0.08			
Photorhabdus			0.02	
Polaromonas	0.15	0.01		
Psychrobacter	0.11			
Ralstonia	0.21	0.07		
Rhizobium	0.07		0.02	
Rhodobacter	0.02			
Rhodopseudomonas	0.05		0.02	
Rhodospirillum	0.03			
Rickettsia			0.05	
Salmonella	0.02	0.11	0.02	0.25
Shewanella	0.03	0.01		
Sinorhizobium	0.02			
Sphingomonas	0.08			
Sphingopyxis	0.05		0.02	
Staphylococcus	0.03		0.03	
Stenotrophomonas	0.64		0.03	
Streptococcus			0.03	
Streptomyces	0.05		0.09	
Thauera	0.03			
Thermobifida			0.04	
Verminephrobacter	0.05			
Vibrio	0.02	0.05		
Wigglesworthia		0.01		
Wolbachia	0.03		0.09	
Xanthomonas	0.16	0.01		
Yersinia		0.07	0.03	

Genus-level taxonomy tree from 454 sequence analysis of Sample 6

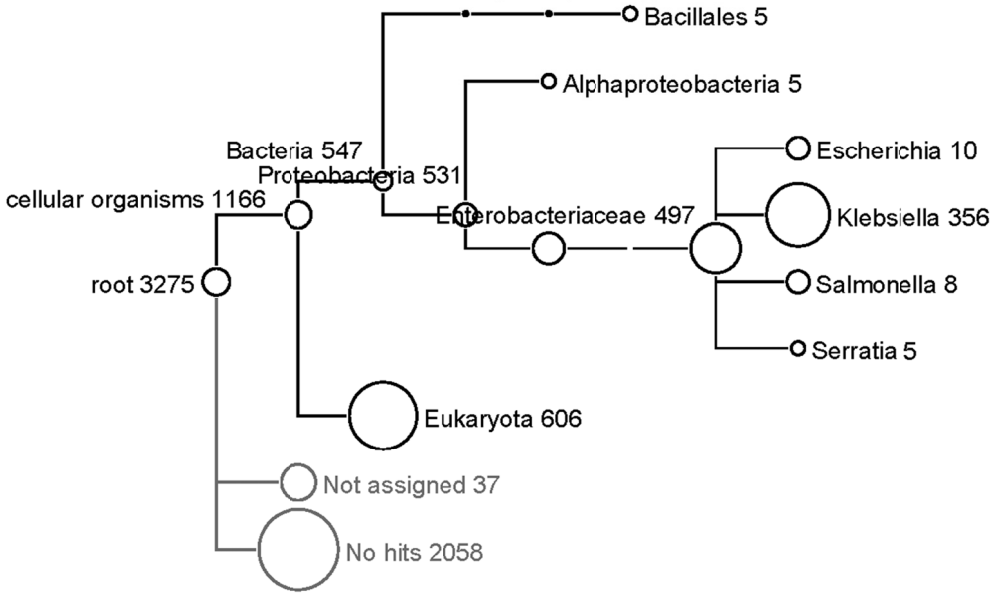


Genus-level taxonomy tree from 454 sequence analysis of Sample 8

Genus-level taxonomy tree from Illumina sequence analysis of Sample 6



Genus-level taxonomy tree from 454 sequence analysis of Sample 8



3.6 Identification of virulence factors or markers

For each metagenomic sample, two distinct analyses based on 454 sequences and Illumina sequences, respectively, were performed to detect the existence of virulence factors or synthetic/genetically modified markers. For each sample, a table is given to show the identified markers and their estimated proportions from the two analyses. The markers are listed in descending order of the maximum (of the two possible) percentage reads.

3.6.1 "Known" group

This group contains Samples 1, 2, 3, and 4.

Markers identified in Sample 1 (35 out of 48)

Identified marker		**454 analysis** (Reads whose BLAST top hit is this marker)		**Illumina analysis** (Reads whose Bowtie best hit is this marker)	
Accession number	**Description**	**Number of reads**	**% reads**	**Number of reads**	**% reads**
gi|386685405|dbj|AB665981.1|	Staphylococcus aureus DNA, ACME and type II SCCmec, complete and partial sequence, strain: SR 388	4,089	1.4722	119,583	0.5687
AY183453	Escherichia coli plasmid pIP1100 class II integron multidrug resistance locus, partial sequence.	1,786	0.6430	49,159	0.2338
X54227	S.epidermidis plasmid pIP1842 fosB gene for FOSB.	216	0.0778		
gi|325965616|gb	Lux single copy cloning vector pMH30	108	0.0389	2,583	0.0123

|JF420885.1|	luxCDABE operon, complete sequence; and aminoglycoside 3'-phosphotransferase (kanR) gene, complete cds				
gi|441494908|gb|KC243783.1|	Staphylococcus aureus strain TN/CN/1/12 MecA (mecA) gene, complete cds	3	0.0011	5,034	0.0239
gi|8926246|gb|AF271719.1|	Clostridium difficile CdtA (cdtA) and CdtB (cdtB) genes, complete cds	49	0.0176		
gi|155036|gb|M97297.1|TRNVAN	Enterococcus faecium transposon Tn1546 transposase, resolvase, vanR (vanR), vanS (vanS), vanH (vanH), vanA (vanA), vanX (vanX), vanY (vanY), and teicoplanin resistance protein (vanZ) genes, complete cds	39	0.0140		
AB113580	Klebsiella pneumoniae integron In112 genes (intI1, blaGES-3, aacA1, orfG, orfA, qacEdelta1), IS26 tnpA, IS6100 tnpA, orf6, orf5 genes, complete and partial cds.	35	0.0126	964	0.0046
M95287	Plasmid R46 class 1 integron In1 integrase, beta-lactamase, aminoglycoside	28	0.0101	390	0.0019
AY887066	Acinetobacter genomosp. 3 isolate YMC 03/9/T104 class I integron DNA integrase (intI1) gene, partial cds, metallo-beta-lactamase SIM-1 (blaSIM-1), ADP-ribosylating transferase (arr-3), chloramphenicol acetyltransferase (catB3), and aminoglycoside 3'-adenyltransferase (aadA1) genes, complete cds, and quaternary ammonium compound-resistance protein (qacEdelta1) gene, partial cds.	24	0.0086	341	0.0016
AF010416	Escherichia coli extended spectrum beta-lactamase (veb-1) gene, complete cds.	9	0.0032	95	0.0005
AF355189	Pseudomonas aeruginosa integron In60 integrase IntI1 (intI1), aminoglycoside 3-N-acetyltransferase/aminoglycoside 6'-N-acetyltransferase fusion protein (aac(3)-Ib/aac(6')-Ib), and beta-lactamase GES-1 (blages-1) genes, complete cds.	7	0.0025	80	0.0004
FJ808975	Acinetobacter johnsonii strain 7037 insertion sequence ISCR2 VEB-1a (blaVEB-1a) and XRE (xre) genes, complete cds, TnpA-like gene, complete sequence, and unknown gene.	7	0.0025		
X12870	Transposon Tn21 (plasmid R6-5) aadA-sulI region for spectinomycin and sulfonamide resistance.	4	0.0014	405	0.0019
U12441.2	Escherichia coli plasmid R388 class 1 integron In3, complete sequence.	3	0.0011	396	0.0019
gi|31321897|gb|AY037297.1|	Synthetic construct erythromycin resistance protein (erm) gene, partial cds; and streptomycin 3'-phosphotransferase (sph), bleomycin phosphotransferase (ble), neomycin	5	0.0018	158	0.0008

	phosphotransferase (nptII), and gentamycin resistance protein (aac) genes, complete cds				
DQ143913	Klebsiella pneumoniae class I integron Inh12 IntI1 (intI1), AacA7 (aacA7), VIM-12 (blaVIM-12), AacA7 (aacA7), QacEdelta1 (qacEdelta1), and Sul1 (sul1) genes, complete cds.			200	0.0010
X12868	Plasmid pLMO20 dhfrV gene for trimethoprim resistance.	2	0.0007	180	0.0009
AJ628983	Pseudomonas aeruginosa class 1 integron, intI1 gene (partial), blaVIM gene, bla gene, qacEdelta1 gene and sul1 gene (partial), strain c6.			160	0.0008
AF347074	Pseudomonas aeruginosa class I integron integrase (intI1) gene, partial cds; expanded-spectrum beta-lactamase GES-2 (bla GES-2), beta-lactamase OXA-5 (bla OXA-5), and aminoglycoside acetyltransferase (aac(3)-I) genes, complete cds; and QACEdelta1 (qacEdelta1) gene, partial cds.	2	0.0007	75	0.0004
gi|15430752|gb|AF403784.1|	Photorhabdus luminescens lux operon, complete sequence	2	0.0007	12	0.0001
AY139601	Uncultured bacterium plasmid pSp39 class I integron gene cassette array dihydrofolate reductase (dfrII), chloramphenicol acetyltransferase (catB2), and aminoglycoside-3'-adenylyltransferase (aadA1) genes, complete cds.	1	0.0004	49	0.0002
AB121039	Escherichia coli estX, aadA2 genes for putative esterase, putative phosphoserine phosphatase, streptomycin-spectinomycin resistance protein, complete cds.	1	0.0004	9	0.0000
gi|145208532|gb|DQ861424.2|	Clostridium difficile variant TcdC (tcdC) gene, tcdC-sc16 allele, complete cds	1	0.0004		
gi|143141|gb|M30210.1|BACLEF	B.anthracis plasmid pX01 lethal factor (lef) gene, complete cds	1	0.0004		
gi|49071|gb|X65462|	S.agalactiae plasmid pIP501 cat gene for chloramphenicol acetyltransferase	1	0.0004		
gi|142630|gb|M24150.1|BACCAPABC	B.anthracis encapsulation protein genes (capA, capB, and capC), complete cds	1	0.0004		
gi|40880|emb|X02340.1|	E. coli R538-1 plasmid aadA gene for aminoglycoside 3 adenyltransferase AAD(3)(9)			77	0.0004
FJ808975	cinetobacter johnsonii strain 7037 insertion sequence ISCR2 VEB-1a (blaVEB-1a) and XRE (xre) genes, complete cds, TnpA-like gene, complete sequence, and unknown gene.			69	0.0003

AY027870	Pseudomonas aeruginosa integron extended-spectrum beta-lactamase VEB-2 (blaVEB-2) gene, complete cds.			61	0.0003
AF324834	Pseudomonas aeruginosa extended spectrum beta-lactamase VEB-1b (bla) gene, complete cds.			51	0.0002
AJ487034	Salmonella enteritidis class 1 integron conatining aacA4 gene, aadA1 gene and catB2 gene.			32	0.0002
AF318077	Pseudomonas aeruginosa integron InAB1 aminoglycoside 6'-N-acetyltransferase (aacC4), metallo-beta-lactamase (blaIMP-7), and aminoglycoside 3-N-acetyltransferase (aacC1) genes, complete cds.			15	0.0001
X15995	E. coli SAT-1 gene for streptothricin-acetyl-transferase.			11	0.0001
JX023441	Acinetobacter baumannii class I integron IntI1 (intI1) gene, partial cds, GES-22 (blaGES-22), aminoglycoside acetyltransferase (6') type I (aac(6')-Ib), and dihydrofolate reductase type A7 (dfrA7) genes, complete cds, and qacEdelta1 (qacEdelta1) gene, partial cds.			1	0.0000

Markers identified in Sample 2 (38 out of 48)

Identified marker		**454 analysis** (Reads whose BLAST top hit is this marker)		**Illumina analysis** (Reads whose Bowtie best hit is this marker)	
Accession number	**Description**	**Number of reads**	**% reads**	**Number of reads**	**% reads**
AY183453	Escherichia coli plasmid pIP1100 class II integron multidrug resistance locus, partial sequence.	4,728	1.7865	221,485	0.7415
gi|325965616|gb|JF420885.1|	Lux single copy cloning vector pMH30 luxCDABE operon, complete sequence; and aminoglycoside 3'-phosphotransferase (kanR) gene, complete cds	297	0.1122	12,232	0.0410
gi|386685405|dbj|AB665981.1|	Staphylococcus aureus DNA, ACME and type II SCCmec, complete and partial sequence, strain: SR 388	173	0.0654	1	0.0000
gi|31321897|gb|AY037297.1|	Synthetic construct erythromycin resistance protein (erm) gene, partial cds; and streptomycin 3'-phosphotransferase (sph), bleomycin phosphotransferase (ble), neomycin phosphotransferase (nptII), and gentamycin resistance protein (aac) genes, complete cds	144	0.0544	6,213	0.0208
gi|155036|gb|M	Enterococcus faecium transposon Tn1546	90	0.0340		

97297.1|TRNVAN	transposase, resolvase, vanR (vanR), vanS (vanS), vanH (vanH), vanA (vanA), vanX (vanX), vanY (vanY), and teicoplanin resistance protein (vanZ) genes, complete cds				
gi|8926246|gb|AF271719.1|	Clostridium difficile CdtA (cdtA) and CdtB (cdtB) genes, complete cds	66	0.0249	4,551	0.0152
gi|118425770|gb|DQ914438.1|	Clostridium difficile strain R11402 toxin A-like (tcdA) gene, partial sequence	55	0.0208	660	0.0022
AB113580	Klebsiella pneumoniae integron In112 genes (intI1, blaGES-3, aacA1, orfG, orfA, qacEdelta1), IS26 tnpA, IS6100 tnpA, orf6, orf5 genes, complete and partial cds.	51	0.0193	2,416	0.0081
M95287	Plasmid R46 class 1 integron In1 integrase, beta-lactamase, aminoglycoside	46	0.0174	1,053	0.0035
gi|145208532|gb|DQ861424.2|	Clostridium difficile variant TcdC (tcdC) gene, tcdC-sc16 allele, complete cds	38	0.0144	1,831	0.0061
gi|15430752|gb|AF403784.1|	Photorhabdus luminescens lux operon, complete sequence	35	0.0132	38	0.0001
gi|143141|gb|M30210.1|BACLEF	B.anthracis plasmid pX01 lethal factor (lef) gene, complete cds	30	0.0113		
AB121039	Escherichia coli estX, aadA2 genes for putative esterase, putative phosphoserine phosphatase, streptomycin-spectinomycin resistance protein, complete cds.	26	0.0098	24	0.0001
AY887066	Acinetobacter genomosp. 3 isolate YMC 03/9/T104 class I integron DNA integrase (intI1) gene, partial cds, metallo-beta-lactamase SIM-1 (blaSIM-1), ADP-ribosylating transferase (arr-3), chloramphenicol acetyltransferase (catB3), and aminoglycoside 3'-adenyltransferase (aadA1) genes, complete cds, and quaternary ammonium compound-resistance protein (qacEdelta1) gene, partial cds.	17	0.0064	907	0.0030
X12870	Transposon Tn21 (plasmid R6-5) aadA-sulI region for spectinomycin and sulfonamide resistance.	16	0.0060	1,017	0.0034
gi|142630|gb|M24150.1|BACCAPABC	B.anthracis encapsulation protein genes (capA, capB, and capC), complete cds	16	0.0060		
FJ808975	Acinetobacter johnsonii strain 7037 insertion sequence ISCR2 VEB-1a (blaVEB-1a) and XRE (xre) genes, complete cds, TnpA-like gene, complete sequence, and unknown gene.	13	0.0049		
AF010416	Escherichia coli extended spectrum beta-lactamase (veb-1) gene, complete cds.	11	0.0042	258	0.0009
U12441.2	Escherichia coli plasmid R388 class 1 integron In3, complete sequence.	1	0.0004	1,114	0.0037

gi|49071|gb|X65462|	S.agalactiae plasmid pIP501 cat gene for chloramphenicol acetyltransferase	6	0.0023		
DQ143913	Klebsiella pneumoniae class I integron Inh12 IntI1 (intI1), AacA7 (aacA7), VIM-12 (blaVIM-12), AacA7 (aacA7), QacEdelta1 (qacEdelta1), and Sul1 (sul1) genes, complete cds.			602	0.0020
X12868	Plasmid pLMO20 dhfrV gene for trimethoprim resistance.	4	0.0015	465	0.0016
AF355189	Pseudomonas aeruginosa integron In60 integrase IntI1 (intI1), aminoglycoside 3-N-acetyltransferase/aminoglycoside 6'-N-acetyltransferase fusion protein (aac(3)-Ib/aac(6')-Ib), and beta-lactamase GES-1 (blages-1) genes, complete cds.	4	0.0015	181	0.0006
X54227	S.epidermidis plasmid pIP1842 fosB gene for FOSB.	4	0.0015		
AJ628983	Pseudomonas aeruginosa class 1 integron, intI1 gene (partial), blaVIM gene, bla gene, qacEdelta1 gene and sul1 gene (partial), strain c6.			445	0.0015
AY139601	Uncultured bacterium plasmid pSp39 class I integron gene cassette array dihydrofolate reductase (dfrII), chloramphenicol acetyltransferase (catB2), and aminoglycoside-3'-adenylyltransferase (aadA1) genes, complete cds.	3	0.0011	114	0.0004
AF347074	Pseudomonas aeruginosa class I integron integrase (intI1) gene, partial cds; expanded-spectrum beta-lactamase GES-2 (bla GES-2), beta-lactamase OXA-5 (bla OXA-5), and aminoglycoside acetyltransferase (aac(3)-I) genes, complete cds; and QACEdelta1 (qacEdelta1) gene, partial cds.	2	0.0008	192	0.0006
gi|40880|emb|X02340.1|	E. coli R538-1 plasmid aadA gene for aminoglycoside 3 adenyltransferase AAD(3)(9)			206	0.0007
FJ808975	cinetobacter johnsonii strain 7037 insertion sequence ISCR2 VEB-1a (blaVEB-1a) and XRE (xre) genes, complete cds, TnpA-like gene, complete sequence, and unknown gene.			169	0.0006
AY027870	Pseudomonas aeruginosa integron extended-spectrum beta-lactamase VEB-2 (blaVEB-2) gene, complete cds.			149	0.0005
AJ487034	Salmonella enteritidis class 1 integron conatining aacA4 gene, aadA1 gene and catB2 gene.	1	0.0004	110	0.0004
AF324834	Pseudomonas aeruginosa extended spectrum beta-lactamase VEB-1b (bla) gene, complete			107	0.0004

	cds.				
gi|47033|emb|V01547.1|	Streptococcus faecalis kanamycin resistance gene encoding a 3'5"-aminoglycoside phosphotransferase of type III. The gene resides on plasmid pJH1	1	0.0004		
AF318077	Pseudomonas aeruginosa integron InAB1 aminoglycoside 6'-N-acetyltransferase (aacC4), metallo-beta-lactamase (blaIMP-7), and aminoglycoside 3-N-acetyltransferase (aacC1) genes, complete cds.			47	0.0002
X15995	E. coli SAT-1 gene for streptothricin-acetyl-transferase.			23	0.0001
gi|281333459|gb|GU211012.1|	Klebsiella pneumoniae strain 2 plasmid SHV-11 beta-lactamase (blaSHV-11) gene, complete cds			2	0.0000
JX023441	Acinetobacter baumannii class I integron IntI1 (intI1) gene, partial cds, GES-22 (blaGES-22), aminoglycoside acetyltransferase (6') type I (aac(6')-Ib), and dihydrofolate reductase type A7 (dfrA7) genes, complete cds, and qacEdelta1 (qacEdelta1) gene, partial cds.			1	0.0000
gi|439981485|gb|JX976326.1|	Escherichia coli plasmid pECDF16 extended spectrum beta-lactamase 2 (TEM-1) gene, complete cds			1	0.0000

Markers identified in Sample 3 (39 out of 48)

Identified marker		**454 analysis** (Reads whose BLAST top hit is this marker)		**Illumina analysis** (Reads whose Bowtie best hit is this marker)	
Accession number	**Description**	**Number of reads**	**% reads**	**Number of reads**	**% reads**
AY183453	Escherichia coli plasmid pIP1100 class II integron multidrug resistance locus, partial sequence.	5,793	2.3934	246,271	1.0973
gi|325965616|gb|JF420885.1|	Lux single copy cloning vector pMH30 luxCDABE operon, complete sequence; and aminoglycoside 3'-phosphotransferase (kanR) gene, complete cds	665	0.2747	13,650	0.0608
AB113580	Klebsiella pneumoniae integron In112 genes (intI1, blaGES-3, aacA1, orfG, orfA, qacEdelta1), IS26 tnpA, IS6100 tnpA, orf6, orf5 genes, complete and partial cds.	303	0.1252	11,398	0.0508
AJ487034	Salmonella enteritidis class 1 integron conatining aacA4 gene, aadA1 gene and catB2 gene.	78	0.0322	1,179	0.0053
gi|281333459|gb	Klebsiella pneumoniae strain 2 plasmid SHV-	75	0.0310	2,223	0.0099

|GU211012.1|	11 beta-lactamase (blaSHV-11) gene, complete cds				
M95287	Plasmid R46 class 1 integron In1 integrase, beta-lactamase, aminoglycoside	51	0.0211	1,199	0.0053
gi|155036|gb|M97297.1|TRNVAN	Enterococcus faecium transposon Tn1546 transposase, resolvase, vanR (vanR), vanS (vanS), vanH (vanH), vanA (vanA), vanX (vanX), vanY (vanY), and teicoplanin resistance protein (vanZ) genes, complete cds	32	0.0132		
gi|439981485|gb|JX976326.1|	Escherichia coli plasmid pECDF16 extended spectrum beta-lactamase 2 (TEM-1) gene, complete cds	27	0.0112	1,094	0.0049
U12441.2	Escherichia coli plasmid R388 class 1 integron In3, complete sequence.	26	0.0107	1,076	0.0048
X12870	Transposon Tn21 (plasmid R6-5) aadA-sulI region for spectinomycin and sulfonamide resistance.	18	0.0074	1,061	0.0047
AF355189	Pseudomonas aeruginosa integron In60 integrase IntI1 (intI1), aminoglycoside 3-N-acetyltransferase/aminoglycoside 6'-N-acetyltransferase fusion protein (aac(3)-Ib/aac(6')-Ib), and beta-lactamase GES-1 (blages-1) genes, complete cds.	15	0.0062	644	0.0029
gi|315940016|gb|HQ162129.1|	Escherichia coli strain MZ28 TEM-2 extended-spectrum beta-lactamase gene, partial cds	15	0.0062	377	0.0017
AB121039	Escherichia coli estX, aadA2 genes for putative esterase, putative phosphoserine phosphatase, streptomycin-spectinomycin resistance protein, complete cds.	14	0.0058	1,230	0.0055
AF318077	Pseudomonas aeruginosa integron InAB1 aminoglycoside 6'-N-acetyltransferase (aacC4), metallo-beta-lactamase (blaIMP-7), and aminoglycoside 3-N-acetyltransferase (aacC1) genes, complete cds.	14	0.0058		
gi|31321897|gb|AY037297.1|	Synthetic construct erythromycin resistance protein (erm) gene, partial cds; and streptomycin 3'-phosphotransferase (sph), bleomycin phosphotransferase (ble), neomycin phosphotransferase (nptII), and gentamycin resistance protein (aac) genes, complete cds	10	0.0041	483	0.0022
gi|386685405|dbj|AB665981.1|	Staphylococcus aureus DNA, ACME and type II SCCmec, complete and partial sequence, strain: SR 388	10	0.0041	3	0.0000
JX023441	Acinetobacter baumannii class I integron IntI1 (intI1) gene, partial cds, GES-22 (blaGES-22), aminoglycoside acetyltransferase (6') type I (aac(6')-Ib), and dihydrofolate reductase type A7 (dfrA7) genes, complete cds, and	1	0.0004	875	0.0039

	qacEdelta1 (qacEdelta1) gene, partial cds.				
gi|262092528|gb|GU086225.1|	Klebsiella pneumoniae strain HZ001 class A carbapenemase blaKPC-2 gene, complete cds	9	0.0037	360	0.0016
gi|224384176|gb|FJ665695.1|	Klebsiella pneumoniae strain ATCC BAA-1705 carbapenem-hydrolyzing beta-lactamase (blaKPC-2) gene, partial cds	8	0.0033	346	0.0015
DQ143913	Klebsiella pneumoniae class I integron Inh12 IntI1 (intI1), AacA7 (aacA7), VIM-12 (blaVIM-12), AacA7 (aacA7), QacEdelta1 (qacEdelta1), and Sul1 (sul1) genes, complete cds.			598	0.0027
AY887066	Acinetobacter genomosp. 3 isolate YMC 03/9/T104 class I integron DNA integrase (intI1) gene, partial cds, metallo-beta-lactamase SIM-1 (blaSIM-1), ADP-ribosylating transferase (arr-3), chloramphenicol acetyltransferase (catB3), and aminoglycoside 3'-adenyltransferase (aadA1) genes, complete cds, and quaternary ammonium compound-resistance protein (qacEdelta1) gene, partial cds.	6	0.0025	495	0.0022
gi|15430752|gb|AF403784.1|	Photorhabdus luminescens lux operon, complete sequence	6	0.0025	45	0.0002
X12868	Plasmid pLMO20 dhfrV gene for trimethoprim resistance.	1	0.0004	476	0.0021
X54227	S.epidermidis plasmid pIP1842 fosB gene for FOSB.	5	0.0021		
AF318077	Pseudomonas aeruginosa integron InAB1 aminoglycoside 6'-N-acetyltransferase (aacC4), metallo-beta-lactamase (blaIMP-7), and aminoglycoside 3-N-acetyltransferase (aacC1) genes, complete cds.			482	0.0021
AJ628983	Pseudomonas aeruginosa class 1 integron, intI1 gene (partial), blaVIM gene, bla gene, qacEdelta1 gene and sul1 gene (partial), strain c6.			448	0.0020
FJ808975	Acinetobacter johnsonii strain 7037 insertion sequence ISCR2 VEB-1a (blaVEB-1a) and XRE (xre) genes, complete cds, TnpA-like gene, complete sequence, and unknown gene.	4	0.0017		
AF010416	Escherichia coli extended spectrum beta-lactamase (veb-1) gene, complete cds.	3	0.0012	51	0.0002
gi|296939581|gb|HM066995.1|	Klebsiella pneumoniae plasmid pKp10-26 beta-lactamase KPC-11 (blaKPC) gene, blaKPC-11 allele, complete cds			279	0.0012
D50438	Serratia marcescens DNA for integrase, metallo-bata-lactamase, aminoglycoside acetyltransferase, complete cds.			248	0.0011
AF347074	Pseudomonas aeruginosa class I integron	2	0.0008	185	0.0008

	integrase (intI1) gene, partial cds; expanded-spectrum beta-lactamase GES-2 (bla GES-2), beta-lactamase OXA-5 (bla OXA-5), and aminoglycoside acetyltransferase (aac(3)-I) genes, complete cds; and QACEdelta1 (qacEdelta1) gene, partial cds.				
gi|8926246|gb|AF271719.1|	Clostridium difficile CdtA (cdtA) and CdtB (cdtB) genes, complete cds	2	0.0008		
gi|40880|emb|X02340.1|	E. coli R538-1 plasmid aadA gene for aminoglycoside 3 adenyltransferase AAD(3)(9)			137	0.0006
AY139601	Uncultured bacterium plasmid pSp39 class I integron gene cassette array dihydrofolate reductase (dfrII), chloramphenicol acetyltransferase (catB2), and aminoglycoside-3'-adenylyltransferase (aadA1) genes, complete cds.			105	0.0005
X15995	E. coli SAT-1 gene for streptothricin-acetyl-transferase.			76	0.0003
FJ808975	cinetobacter johnsonii strain 7037 insertion sequence ISCR2 VEB-1a (blaVEB-1a) and XRE (xre) genes, complete cds, TnpA-like gene, complete sequence, and unknown gene.			27	0.0001
AY027870	Pseudomonas aeruginosa integron extended-spectrum beta-lactamase VEB-2 (blaVEB-2) gene, complete cds.			23	0.0001
AF324834	Pseudomonas aeruginosa extended spectrum beta-lactamase VEB-1b (bla) gene, complete cds.			20	0.0001
gi|118425770|gb|DQ914438.1|	Clostridium difficile strain R11402 toxin A-like (tcdA) gene, partial sequence			1	0.0000

Markers identified in Sample 4 (17 out of 48)

Identified marker		**454 analysis** (Reads whose BLAST top hit is this marker)		**Illumina analysis** (Reads whose Bowtie best hit is this marker)	
Accession number	**Description**	**Number of reads**	**% reads**	**Number of reads**	**% reads**
gi|325965616|gb|JF420885.1|	Lux single copy cloning vector pMH30 luxCDABE operon, complete sequence; and aminoglycoside 3'-phosphotransferase (kanR) gene, complete cds	757	0.2117	18,437	0.0919
gi|386685405|dbj|AB665981.1|	Staphylococcus aureus DNA, ACME and type II SCCmec, complete and partial sequence, strain: SR 388	550	0.1538	774	0.0039
X54227	S.epidermidis plasmid pIP1842 fosB gene for	113	0.0316	154	0.0008

	FOSB.				
gi|15430752|gb|AF403784.1|	Photorhabdus luminescens lux operon, complete sequence	90	0.0252	60	0.0003
gi|31321897|gb|AY037297.1|	Synthetic construct erythromycin resistance protein (erm) gene, partial cds; and streptomycin 3'-phosphotransferase (sph), bleomycin phosphotransferase (ble), neomycin phosphotransferase (nptII), and gentamycin resistance protein (aac) genes, complete cds	44	0.0123		
gi|47033|emb|V01547.1|	Streptococcus faecalis kanamycin resistance gene encoding a 3'5"-aminoglycoside phosphotransferase of type III. The gene resides on plasmid pJH1	35	0.0098	1,172	0.0058
AY887066	Acinetobacter genomosp. 3 isolate YMC 03/9/T104 class I integron DNA integrase (intI1) gene, partial cds, metallo-beta-lactamase SIM-1 (blaSIM-1), ADP-ribosylating transferase (arr-3), chloramphenicol acetyltransferase (catB3), and aminoglycoside 3'-adenyltransferase (aadA1) genes, complete cds, and quaternary ammonium compound-resistance protein (qacEdelta1) gene, partial cds.	29	0.0081		
AB113580	Klebsiella pneumoniae integron In112 genes (intI1, blaGES-3, aacA1, orfG, orfA, qacEdelta1), IS26 tnpA, IS6100 tnpA, orf6, orf5 genes, complete and partial cds.	23	0.0064	5	0.0000
gi|8926246|gb|AF271719.1|	Clostridium difficile CdtA (cdtA) and CdtB (cdtB) genes, complete cds	20	0.0056		
gi|143141|gb|M30210.1|BACLEF	B.anthracis plasmid pX01 lethal factor (lef) gene, complete cds	11	0.0031	5	0.0000
AY183453	Escherichia coli plasmid pIP1100 class II integron multidrug resistance locus, partial sequence.	3	0.0008	78	0.0004
gi|142630|gb|M24150.1|BACCAPABC	B.anthracis encapsulation protein genes (capA, capB, and capC), complete cds	2	0.0006	1	0.0000
gi|118425770|gb|DQ914438.1|	Clostridium difficile strain R11402 toxin A-like (tcdA) gene, partial sequence	1	0.0003		
gi|441494908|gb|KC243783.1|	Staphylococcus aureus strain TN/CN/1/12 MecA (mecA) gene, complete cds			2	0.0000
gi|29786397|gb|AJ413935.2|	Bacillus anthracis partial lef gene, isolate IT-Carb3-6259.			2	0.0000
M95287	Plasmid R46 class 1 integron In1 integrase, beta-lactamase, aminoglycoside			1	0.0000
gi|281333459|gb|GU211012.1|	Klebsiella pneumoniae strain 2 plasmid SHV-11 beta-lactamase (blaSHV-11) gene, complete cds			1	0.0000

3.6.2 "Partially known" group

This group contains Samples 7 and 14.

Markers identified in Sample 7 (22 out of 48)

Identified marker		**454 analysis** (Reads whose BLAST top hit is this marker)		**Illumina analysis** (Reads whose Bowtie best hit is this marker)	
Accession number	**Description**	**Number of reads**	**% reads**	**Number of reads**	**% reads**
gi|325965616|gb|JF420885.1|	Lux single copy cloning vector pMH30 luxCDABE operon, complete sequence; and aminoglycoside 3'-phosphotransferase (kanR) gene, complete cds	1,602	0.5723	98,240	0.3183
gi|31321897|gb|AY037297.1|	Synthetic construct erythromycin resistance protein (erm) gene, partial cds; and streptomycin 3'-phosphotransferase (sph), bleomycin phosphotransferase (ble), neomycin phosphotransferase (nptII), and gentamycin resistance protein (aac) genes, complete cds	127	0.0454	1	0.0000
gi|15430752|gb|AF403784.1|	Photorhabdus luminescens lux operon, complete sequence	93	0.0332	311	0.0010
AB113580	Klebsiella pneumoniae integron In112 genes (intI1, blaGES-3, aacA1, orfG, orfA, qacEdelta1), IS26 tnpA, IS6100 tnpA, orf6, orf5 genes, complete and partial cds.	59	0.0211	1	0.0000
gi|143141|gb|M30210.1|BACLEF	B.anthracis plasmid pX01 lethal factor (lef) gene, complete cds	37	0.0132		
X54227	S.epidermidis plasmid pIP1842 fosB gene for FOSB.	10	0.0036		
AY887066	Acinetobacter genomosp. 3 isolate YMC 03/9/T104 class I integron DNA integrase (intI1) gene, partial cds, metallo-beta-lactamase SIM-1 (blaSIM-1), ADP-ribosylating transferase (arr-3), chloramphenicol acetyltransferase (catB3), and aminoglycoside 3'-adenyltransferase (aadA1) genes, complete cds, and quaternary ammonium compound-resistance protein (qacEdelta1) gene, partial cds.	4	0.0014		
gi|386685405|dbj|AB665981.1|	Staphylococcus aureus DNA, ACME and type II SCCmec, complete and partial sequence, strain: SR 388	3	0.0011	12	0.0000
M95287	Plasmid R46 class 1 integron In1 integrase, beta-lactamase, aminoglycoside	2	0.0007		
gi|439981485|gb	Escherichia coli plasmid pECDF16 extended	1	0.0004	1	0.0000

|JX976326.1|	spectrum beta-lactamase 2 (TEM-1) gene, complete cds				
gi|444746673|gb|KC347597.1|	Acinetobacter baumannii strain B2214 metallo-beta-lactamase-1 (blaNDM-1) gene, complete cds	1	0.0004		
gi|155036|gb|M97297.1|TRNVAN	Enterococcus faecium transposon Tn1546 transposase, resolvase, vanR (vanR), vanS (vanS), vanH (vanH), vanA (vanA), vanX (vanX), vanY (vanY), and teicoplanin resistance protein (vanZ) genes, complete cds	1	0.0004		
gi|281333459|gb|GU211012.1|	Klebsiella pneumoniae strain 2 plasmid SHV-11 beta-lactamase (blaSHV-11) gene, complete cds	1	0.0004		
AJ628983	Pseudomonas aeruginosa class 1 integron, intI1 gene (partial), blaVIM gene, bla gene, qacEdelta1 gene and sul1 gene (partial), strain c6.	1	0.0004		
FJ808975	Acinetobacter johnsonii strain 7037 insertion sequence ISCR2 VEB-1a (blaVEB-1a) and XRE (xre) genes, complete cds, TnpA-like gene, complete sequence, and unknown gene.	1	0.0004		
AF355189	Pseudomonas aeruginosa integron In60 integrase IntI1 (intI1), aminoglycoside 3-N-acetyltransferase/aminoglycoside 6'-N-acetyltransferase fusion protein (aac(3)-Ib/aac(6')-Ib), and beta-lactamase GES-1 (blages-1) genes, complete cds.	1	0.0004		
D50438	Serratia marcescens DNA for integrase, metallo-bata-lactamase, aminoglycoside acetyltransferase, complete cds.	1	0.0004		
gi|224384176|gb|FJ665695.1|	Klebsiella pneumoniae strain ATCC BAA-1705 carbapenem-hydrolyzing beta-lactamase (blaKPC-2) gene, partial cds	1	0.0004		
AY183453	Escherichia coli plasmid pIP1100 class II integron multidrug resistance locus, partial sequence.			46	0.0001
AF347074	Pseudomonas aeruginosa class I integron integrase (intI1) gene, partial cds; expanded-spectrum beta-lactamase GES-2 (bla GES-2), beta-lactamase OXA-5 (bla OXA-5), and aminoglycoside acetyltransferase (aac(3)-I) genes, complete cds; and QACEdelta1 (qacEdelta1) gene, partial cds.			2	0.0000
U12441.2	Escherichia coli plasmid R388 class 1 integron In3, complete sequence.			1	0.0000
X12868	Plasmid pLMO20 dhfrV gene for trimethoprim resistance.			1	0.0000

Markers identified in Sample 14 (34 out of 48)

Identified marker		**454 analysis** (Reads whose BLAST top hit is this marker)		**Illumina analysis** (Reads whose Bowtie best hit is this marker)	
Accession number	**Description**	**Number of reads**	**% reads**	**Number of reads**	**% reads**
AY183453	Escherichia coli plasmid pIP1100 class II integron multidrug resistance locus, partial sequence.	1,124	0.3375	27,372	0.0893
gi|325965616|gb|JF420885.1|	Lux single copy cloning vector pMH30 luxCDABE operon, complete sequence; and aminoglycoside 3'-phosphotransferase (kanR) gene, complete cds	719	0.2159	27,125	0.0885
AB113580	Klebsiella pneumoniae integron In112 genes (intI1, blaGES-3, aacA1, orfG, orfA, qacEdelta1), IS26 tnpA, IS6100 tnpA, orf6, orf5 genes, complete and partial cds.	524	0.1573	27,809	0.0907
M95287	Plasmid R46 class 1 integron In1 integrase, beta-lactamase, aminoglycoside	399	0.1198	6,381	0.0208
gi|439981485|gb|JX976326.1|	Escherichia coli plasmid pECDF16 extended spectrum beta-lactamase 2 (TEM-1) gene, complete cds	113	0.0339	5,132	0.0167
gi|281333459|gb|GU211012.1|	Klebsiella pneumoniae strain 2 plasmid SHV-11 beta-lactamase (blaSHV-11) gene, complete cds	102	0.0306	3,250	0.0106
X12868	Plasmid pLMO20 dhfrV gene for trimethoprim resistance.	97	0.0291	2,599	0.0085
gi|315940016|gb|HQ162129.1|	Escherichia coli strain MZ28 TEM-2 extended-spectrum beta-lactamase gene, partial cds	74	0.0222	2,300	0.0075
AF355189	Pseudomonas aeruginosa integron In60 integrase IntI1 (intI1), aminoglycoside 3-N-acetyltransferase/aminoglycoside 6'-N-acetyltransferase fusion protein (aac(3)-Ib/aac(6')-Ib), and beta-lactamase GES-1 (blages-1) genes, complete cds.	55	0.0165	1,576	0.0051
gi|155036|gb|M97297.1|TRNVAN	Enterococcus faecium transposon Tn1546 transposase, resolvase, vanR (vanR), vanS (vanS), vanH (vanH), vanA (vanA), vanX (vanX), vanY (vanY), and teicoplanin resistance protein (vanZ) genes, complete cds	42	0.0126	7	0.0000
gi|15430752|gb|AF403784.1|	Photorhabdus luminescens lux operon, complete sequence	38	0.0114	86	0.0003
gi|386685405|dbj|AB665981.1|	Staphylococcus aureus DNA, ACME and type II SCCmec, complete and partial sequence, strain: SR 388	20	0.0060	50	0.0002
U12441.2	Escherichia coli plasmid R388 class 1 integron	2	0.0006	1,366	0.0045

	In3, complete sequence.				
gi|31321897|gb|AY037297.1|	Synthetic construct erythromycin resistance protein (erm) gene, partial cds; and streptomycin 3'-phosphotransferase (sph), bleomycin phosphotransferase (ble), neomycin phosphotransferase (nptII), and gentamycin resistance protein (aac) genes, complete cds	12	0.0036	421	0.0014
X12870	Transposon Tn21 (plasmid R6-5) aadA-sulI region for spectinomycin and sulfonamide resistance.			986	0.0032
DQ143913	Klebsiella pneumoniae class I integron Inh12 IntI1 (intI1), AacA7 (aacA7), VIM-12 (blaVIM-12), AacA7 (aacA7), QacEdelta1 (qacEdelta1), and Sul1 (sul1) genes, complete cds.			825	0.0027
AJ628983	Pseudomonas aeruginosa class 1 integron, intI1 gene (partial), blaVIM gene, bla gene, qacEdelta1 gene and sul1 gene (partial), strain c6.			679	0.0022
AF347074	Pseudomonas aeruginosa class I integron integrase (intI1) gene, partial cds; expanded-spectrum beta-lactamase GES-2 (bla GES-2), beta-lactamase OXA-5 (bla OXA-5), and aminoglycoside acetyltransferase (aac(3)-I) genes, complete cds; and QACEdelta1 (qacEdelta1) gene, partial cds.	4	0.0012	279	0.0009
AF318077	Pseudomonas aeruginosa integron InAB1 aminoglycoside 6'-N-acetyltransferase (aacC4), metallo-beta-lactamase (blaIMP-7), and aminoglycoside 3-N-acetyltransferase (aacC1) genes, complete cds.	3	0.0009		
X54227	S.epidermidis plasmid pIP1842 fosB gene for FOSB.	3	0.0009		
gi|47033|emb|V01547.1|	Streptococcus faecalis kanamycin resistance gene encoding a 3'5"-aminoglycoside phosphotransferase of type III. The gene resides on plasmid pJH1	3	0.0009		
AY887066	Acinetobacter genomosp. 3 isolate YMC 03/9/T104 class I integron DNA integrase (intI1) gene, partial cds, metallo-beta-lactamase SIM-1 (blaSIM-1), ADP-ribosylating transferase (arr-3), chloramphenicol acetyltransferase (catB3), and aminoglycoside 3'-adenyltransferase (aadA1) genes, complete cds, and quaternary ammonium compound-resistance protein (qacEdelta1) gene, partial cds.			251	0.0008
AF010416	Escherichia coli extended spectrum beta-lactamase (veb-1) gene, complete cds.	2	0.0006		

AJ487034	Salmonella enteritidis class 1 integron conatining aacA4 gene, aadA1 gene and catB2 gene.	2	0.0006		
gi|224384176|gb|FJ665695.1|	Klebsiella pneumoniae strain ATCC BAA-1705 carbapenem-hydrolyzing beta-lactamase (blaKPC-2) gene, partial cds	2	0.0006		
D50438	Serratia marcescens DNA for integrase, metallo-bata-lactamase, aminoglycoside acetyltransferase, complete cds.	2	0.0006		
gi|142630|gb|M24150.1|BACCAPABC	B.anthracis encapsulation protein genes (capA, capB, and capC), complete cds	2	0.0006		
gi|145208532|gb|DQ861424.2|	Clostridium difficile variant TcdC (tcdC) gene, tcdC-sc16 allele, complete cds	1	0.0003		
gi|143141|gb|M30210.1|BACLEF	B.anthracis plasmid pX01 lethal factor (lef) gene, complete cds	1	0.0003		
AB121039	Escherichia coli estX, aadA2 genes for putative esterase, putative phosphoserine phosphatase, streptomycin-spectinomycin resistance protein, complete cds.	1	0.0003		
gi|40880|emb|X02340.1|	E. coli R538-1 plasmid aadA gene for aminoglycoside 3 adenyltransferase AAD(3)(9)			27	0.0001
AF318077	Pseudomonas aeruginosa integron InAB1 aminoglycoside 6'-N-acetyltransferase (aacC4), metallo-beta-lactamase (blaIMP-7), and aminoglycoside 3-N-acetyltransferase (aacC1) genes, complete cds.			4	0.0000
AY139601	Uncultured bacterium plasmid pSp39 class I integron gene cassette array dihydrofolate reductase (dfrII), chloramphenicol acetyltransferase (catB2), and aminoglycoside-3'-adenylyltransferase (aadA1) genes, complete cds.			2	0.0000
D50438	Serratia marcescens DNA for integrase, metallo-bata-lactamase, aminoglycoside acetyltransferase, complete cds.			1	0.0000

3.6.3 "Unknown" group

This group consists of Samples 6 and 8.

Markers identified in Sample 6 (33 out of 48)

Identified marker		**454 analysis** (Reads whose BLAST top hit is this marker)		**Illumina analysis** (Reads whose Bowtie best hit is this marker)	
Accession number	**Description**	**Number of reads**	**% reads**	**Number of reads**	**% reads**
gi|386685405|dbj|AB665981.1|	Staphylococcus aureus DNA, ACME and type II SCCmec, complete and partial sequence, strain: SR 388	71	0.0260	21	0.0001
gi|8926246|gb|AF271719.1|	Clostridium difficile CdtA (cdtA) and CdtB (cdtB) genes, complete cds	16	0.0059	1	0.0000
gi|325965616|gb|JF420885.1|	Lux single copy cloning vector pMH30 luxCDABE operon, complete sequence; and aminoglycoside 3'-phosphotransferase (kanR) gene, complete cds	16	0.0059	34	0.0001
gi|155036|gb|M97297.1|TRNVAN	Enterococcus faecium transposon Tn1546 transposase, resolvase, vanR (vanR), vanS (vanS), vanH (vanH), vanA (vanA), vanX (vanX), vanY (vanY), and teicoplanin resistance protein (vanZ) genes, complete cds	11	0.0040		
AF347074	Pseudomonas aeruginosa class I integron integrase (intI1) gene, partial cds; expanded-spectrum beta-lactamase GES-2 (bla GES-2), beta-lactamase OXA-5 (bla OXA-5), and aminoglycoside acetyltransferase (aac(3)-I) genes, complete cds; and QACEdelta1 (qacEdelta1) gene, partial cds.	10	0.0037	1	0.0000
gi|31321897|gb|AY037297.1|	Synthetic construct erythromycin resistance protein (erm) gene, partial cds; and streptomycin 3'-phosphotransferase (sph), bleomycin phosphotransferase (ble), neomycin phosphotransferase (nptII), and gentamycin resistance protein (aac) genes, complete cds	9	0.0033	2	0.0000
X54227	S.epidermidis plasmid pIP1842 fosB gene for FOSB.	9	0.0033	2	0.0000
gi|15430752|gb|AF403784.1|	Photorhabdus luminescens lux operon, complete sequence	8	0.0029		
gi|281333459|gb|GU211012.1|	Klebsiella pneumoniae strain 2 plasmid SHV-11 beta-lactamase (blaSHV-11) gene, complete cds	8	0.0029		
gi|143141|gb|M30210.1|BACLEF	B.anthracis plasmid pX01 lethal factor (lef) gene, complete cds	7	0.0026		

gi|142630|gb|M24150.1|BACCAPABC	B.anthracis encapsulation protein genes (capA, capB, and capC), complete cds	7	0.0026		
AY183453	Escherichia coli plasmid pIP1100 class II integron multidrug resistance locus, partial sequence.	5	0.0018	54	0.0002
AB113580	Klebsiella pneumoniae integron In112 genes (intI1, blaGES-3, aacA1, orfG, orfA, qacEdelta1), IS26 tnpA, IS6100 tnpA, orf6, orf5 genes, complete and partial cds.	5	0.0018	11	0.0000
AF318077	Pseudomonas aeruginosa integron InAB1 aminoglycoside 6'-N-acetyltransferase (aacC4), metallo-beta-lactamase (blaIMP-7), and aminoglycoside 3-N-acetyltransferase (aacC1) genes, complete cds.	4	0.0015		
D50438	Serratia marcescens DNA for integrase, metallo-bata-lactamase, aminoglycoside acetyltransferase, complete cds.	4	0.0015		
M95287	Plasmid R46 class 1 integron In1 integrase, beta-lactamase, aminoglycoside	2	0.0007	2	0.0000
AJ628983	Pseudomonas aeruginosa class 1 integron, intI1 gene (partial), blaVIM gene, bla gene, qacEdelta1 gene and sul1 gene (partial), strain c6.	2	0.0007		
JX023441	Acinetobacter baumannii class I integron IntI1 (intI1) gene, partial cds, GES-22 (blaGES-22), aminoglycoside acetyltransferase (6') type I (aac(6')-Ib), and dihydrofolate reductase type A7 (dfrA7) genes, complete cds, and qacEdelta1 (qacEdelta1) gene, partial cds.	2	0.0007		
X12868	Plasmid pLMO20 dhfrV gene for trimethoprim resistance.	1	0.0004	2	0.0000
gi|145208532|gb|DQ861424.2|	Clostridium difficile variant TcdC (tcdC) gene, tcdC-sc16 allele, complete cds	1	0.0004	1	0.0000
AF010416	Escherichia coli extended spectrum beta-lactamase (veb-1) gene, complete cds.	1	0.0004		
AY887066	Acinetobacter genomosp. 3 isolate YMC 03/9/T104 class I integron DNA integrase (intI1) gene, partial cds, metallo-beta-lactamase SIM-1 (blaSIM-1), ADP-ribosylating transferase (arr-3), chloramphenicol acetyltransferase (catB3), and aminoglycoside 3'-adenyltransferase (aadA1) genes, complete cds, and quaternary ammonium compound-resistance protein (qacEdelta1) gene, partial cds.	1	0.0004		
gi|49071|gb|X65462|	S.agalactiae plasmid pIP501 cat gene for chloramphenicol acetyltransferase	1	0.0004		
gi|1244503|gb|U	Enterococcus faecalis insertion sequence IS16	1	0.0004		

35366.1|EFU35366	putative transposase (tnp) gene vanB, complete cds				
DQ143913	Klebsiella pneumoniae class I integron Inh12 IntI1 (intI1), AacA7 (aacA7), VIM-12 (blaVIM-12), AacA7 (aacA7), QacEdelta1 (qacEdelta1), and Sul1 (sul1) genes, complete cds.	1	0.0004		
gi|118425770|gb|DQ914438.1|	Clostridium difficile strain R11402 toxin A-like (tcdA) gene, partial sequence	1	0.0004		
AF324834	Pseudomonas aeruginosa extended spectrum beta-lactamase VEB-1b (bla) gene, complete cds.	1	0.0004		
AF099140	Providencia stuartii plasmid pLQ1723 integron erythromycin esterase (ereA2) gene, complete cds.	1	0.0004		
FJ808975	Acinetobacter johnsonii strain 7037 insertion sequence ISCR2 VEB-1a (blaVEB-1a) and XRE (xre) genes, complete cds, TnpA-like gene, complete sequence, and unknown gene.	1	0.0004		
AY139601	Uncultured bacterium plasmid pSp39 class I integron gene cassette array dihydrofolate reductase (dfrII), chloramphenicol acetyltransferase (catB2), and aminoglycoside-3'-adenylyltransferase (aadA1) genes, complete cds.	1	0.0004		
AF355189	Pseudomonas aeruginosa integron In60 integrase IntI1 (intI1), aminoglycoside 3-N-acetyltransferase/aminoglycoside 6'-N-acetyltransferase fusion protein (aac(3)-Ib/aac(6')-Ib), and beta-lactamase GES-1 (blages-1) genes, complete cds.	1	0.0004		
gi|47033|emb|V01547.1|	Streptococcus faecalis kanamycin resistance gene encoding a 3'5"-aminoglycoside phosphotransferase of type III. The gene resides on plasmid pJH1			2	0.0000
U12441.2	Escherichia coli plasmid R388 class 1 integron In3, complete sequence.			1	0.0000

Markers identified in Sample 8 (37 out of 48)

Identified marker		**454 analysis** (Reads whose BLAST top hit is this marker)		**Illumina analysis** (Reads whose Bowtie best hit is this marker)	
Accession number	**Description**	**Number of reads**	**% reads**	**Number of reads**	**% reads**
gi|386685405|dbj|AB665981.1|	Staphylococcus aureus DNA, ACME and type II SCCmec, complete and partial sequence, strain: SR 388	86	0.0366	18	0.0001
gi|325965616|gb|JF420885.1|	Lux single copy cloning vector pMH30 luxCDABE operon, complete sequence; and aminoglycoside 3'-phosphotransferase (kanR) gene, complete cds	37	0.0157	3,513	0.0139
AB113580	Klebsiella pneumoniae integron In112 genes (intI1, blaGES-3, aacA1, orfG, orfA, qacEdelta1), IS26 tnpA, IS6100 tnpA, orf6, orf5 genes, complete and partial cds.	11	0.0047	3,368	0.0134
AY183453	Escherichia coli plasmid pIP1100 class II integron multidrug resistance locus, partial sequence.	3	0.0013	3,261	0.0129
gi|15430752|gb|AF403784.1|	Photorhabdus luminescens lux operon, complete sequence	13	0.0055	9	0.0000
gi|8926246|gb|AF271719.1|	Clostridium difficile CdtA (cdtA) and CdtB (cdtB) genes, complete cds	12	0.0051	1	0.0000
gi|143141|gb|M30210.1|BACLEF	B.anthracis plasmid pX01 lethal factor (lef) gene, complete cds	10	0.0043		
gi|155036|gb|M97297.1|TRNVAN	Enterococcus faecium transposon Tn1546 transposase, resolvase, vanR (vanR), vanS (vanS), vanH (vanH), vanA (vanA), vanX (vanX), vanY (vanY), and teicoplanin resistance protein (vanZ) genes, complete cds	10	0.0043	2	0.0000
M95287	Plasmid R46 class 1 integron In1 integrase, beta-lactamase, aminoglycoside	1	0.0004	781	0.0031
gi|439981485|gb|JX976326.1|	Escherichia coli plasmid pECDF16 extended spectrum beta-lactamase 2 (TEM-1) gene, complete cds	4	0.0017	674	0.0027
gi|31321897|gb|AY037297.1|	Synthetic construct erythromycin resistance protein (erm) gene, partial cds; and streptomycin 3'-phosphotransferase (sph), bleomycin phosphotransferase (ble), neomycin phosphotransferase (nptII), and gentamycin resistance protein (aac) genes, complete cds	6	0.0026	40	0.0002
gi|118425770|gb|DQ914438.1|	Clostridium difficile strain R11402 toxin A-like (tcdA) gene, partial sequence	5	0.0021		
AF010416	Escherichia coli extended spectrum beta-lactamase (veb-1) gene, complete cds.	4	0.0017		

gi|281333459|gb|GU211012.1|	Klebsiella pneumoniae strain 2 plasmid SHV-11 beta-lactamase (blaSHV-11) gene, complete cds			349	0.0014
X54227	S.epidermidis plasmid pIP1842 fosB gene for FOSB.	3	0.0013	3	0.0000
gi|145208532|gb|DQ861424.2|	Clostridium difficile variant TcdC (tcdC) gene, tcdC-sc16 allele, complete cds	3	0.0013	1	0.0000
gi|47033|emb|V01547.1|	Streptococcus faecalis kanamycin resistance gene encoding a 3'5"-aminoglycoside phosphotransferase of type III. The gene resides on plasmid pJH1	3	0.0013	6	0.0000
X12868	Plasmid pLMO20 dhfrV gene for trimethoprim resistance.			307	0.0012
gi|315940016|gb|HQ162129.1|	Escherichia coli strain MZ28 TEM-2 extended-spectrum beta-lactamase gene, partial cds			282	0.0011
gi|49071|gb|X65462|	S.agalactiae plasmid pIP501 cat gene for chloramphenicol acetyltransferase	2	0.0009		
AJ487034	Salmonella enteritidis class 1 integron conatining aacA4 gene, aadA1 gene and catB2 gene.	2	0.0009		
AF355189	Pseudomonas aeruginosa integron In60 integrase IntI1 (intI1), aminoglycoside 3-N-acetyltransferase/aminoglycoside 6'-N-acetyltransferase fusion protein (aac(3)-Ib/aac(6')-Ib), and beta-lactamase GES-1 (blages-1) genes, complete cds.			168	0.0007
U12441.2	Escherichia coli plasmid R388 class 1 integron In3, complete sequence.			155	0.0006
X12870	Transposon Tn21 (plasmid R6-5) aadA-sulI region for spectinomycin and sulfonamide resistance.			137	0.0005
DQ143913	Klebsiella pneumoniae class I integron Inh12 IntI1 (intI1), AacA7 (aacA7), VIM-12 (blaVIM-12), AacA7 (aacA7), QacEdelta1 (qacEdelta1), and Sul1 (sul1) genes, complete cds.			109	0.0004
AF318077	Pseudomonas aeruginosa integron InAB1 aminoglycoside 6'-N-acetyltransferase (aacC4), metallo-beta-lactamase (blaIMP-7), and aminoglycoside 3-N-acetyltransferase (aacC1) genes, complete cds.	1	0.0004		
AB121039	Escherichia coli estX, aadA2 genes for putative esterase, putative phosphoserine phosphatase, streptomycin-spectinomycin resistance protein, complete cds.	1	0.0004		
FJ808975	Acinetobacter johnsonii strain 7037 insertion sequence ISCR2 VEB-1a (blaVEB-1a) and XRE (xre) genes, complete cds, TnpA-like	1	0.0004		

	gene, complete sequence, and unknown gene.				
gi|142630|gb|M24150.1|BACCAPABC	B.anthracis encapsulation protein genes (capA, capB, and capC), complete cds	1	0.0004		
AJ628983	Pseudomonas aeruginosa class 1 integron, intI1 gene (partial), blaVIM gene, bla gene, qacEdelta1 gene and sul1 gene (partial), strain c6.			87	0.0003
AY887066	Acinetobacter genomosp. 3 isolate YMC 03/9/T104 class I integron DNA integrase (intI1) gene, partial cds, metallo-beta-lactamase SIM-1 (blaSIM-1), ADP-ribosylating transferase (arr-3), chloramphenicol acetyltransferase (catB3), and aminoglycoside 3'-adenyltransferase (aadA1) genes, complete cds, and quaternary ammonium compound-resistance protein (qacEdelta1) gene, partial cds.			27	0.0001
AF347074	Pseudomonas aeruginosa class I integron integrase (intI1) gene, partial cds; expanded-spectrum beta-lactamase GES-2 (bla GES-2), beta-lactamase OXA-5 (bla OXA-5), and aminoglycoside acetyltransferase (aac(3)-I) genes, complete cds; and QACEdelta1 (qacEdelta1) gene, partial cds.			25	0.0001
gi|40880|emb|X02340.1|	E. coli R538-1 plasmid aadA gene for aminoglycoside 3 adenyltransferase AAD(3)(9)			4	0.0000
gi|224384176|gb|FJ665695.1|	Klebsiella pneumoniae strain ATCC BAA-1705 carbapenem-hydrolyzing beta-lactamase (blaKPC-2) gene, partial cds			1	0.0000
gi|441494908|gb|KC243783.1|	Staphylococcus aureus strain TN/CN/1/12 MecA (mecA) gene, complete cds			1	0.0000
AF318077	Pseudomonas aeruginosa integron InAB1 aminoglycoside 6'-N-acetyltransferase (aacC4), metallo-beta-lactamase (blaIMP-7), and aminoglycoside 3-N-acetyltransferase (aacC1) genes, complete cds.			1	0.0000
AY139601	Uncultured bacterium plasmid pSp39 class I integron gene cassette array dihydrofolate reductase (dfrII), chloramphenicol acetyltransferase (catB2), and aminoglycoside-3'-adenylyltransferase (aadA1) genes, complete cds.			1	0.0000

4 Analysis of single-organism sequences

4.1 Summary

Objective

The last experiment in this project will be to test if a sample can be processed using only short read sequence to identify the synthetic or atypical gene sequences from an already sequenced organism.

Conclusion

We were able to identify 10 makers from the Illumina sequences of a single-organism sample. The Illumina sequences have been first mapped to the reference genome (using Bowtie) and then to the marker sequences (using both Bowtie and BLAST).

4.2 Processing of Illumina sequences

Illumina sequences of one sample (Sample 5), which contains a single genetically modified organism, have been analyzed. The test organism is *Yersinia pseudotuberculosis*, for which complete genome sequence of the unmodified organism is available (http://www.genome.jp/dbget-bin/www_bget?refseq+NC_010465). The goal of this analysis was twofold:

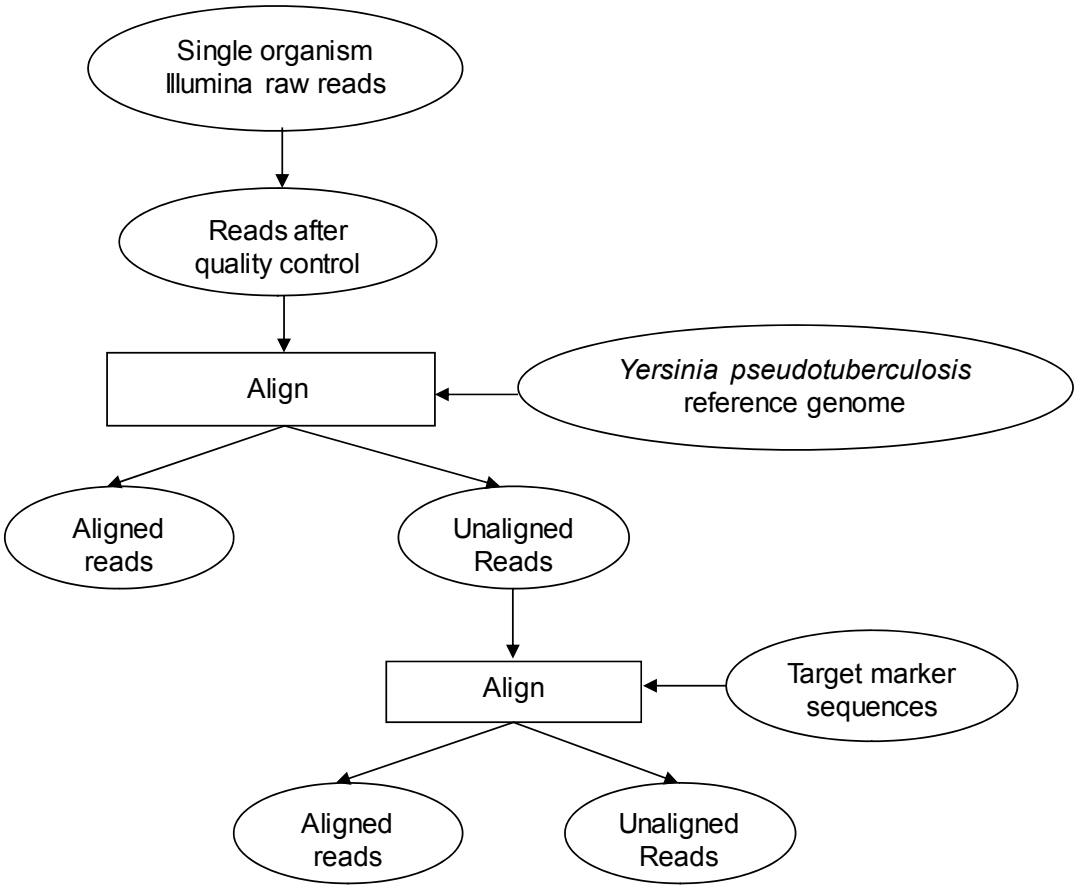
- Detect genetically modified sequences, which are absent in the genome of the unmodified organism
- Among the genetically modified sequences thus found, identify virulence factors or synthetic / genetically modified markers based on the target mark sequences. The targets are the same as the ones used for marker detection using metagenomic sequences (48 sequences).

Although paired-end reads were available for this sample, the R2 reads were of poor quality. Their read lengths became too short after quality control to perform reliable alignment without assembly. Therefore, only R1 reads were used for this analysis. The quality control parameters for the R1 reads were set as follows:

- Trim from the end until the quality score of the last base is at least 20
- Filter if the read length (after trimming) is less than 50 for R1, there is an ambiguous (N) base, or the average quality score over all bases is less than 27

The workflow of this analysis consists of three major stages, as shown in the diagram below:

- Quality control of the raw reads
- Alignment of the quality-controlled reads to the reference genome
- Alignment of the reference-unaligned reads to the target marker sequences



4.3 Alignment to reference genome and to maker sequences

The initial alignment to the reference genome was performed using the Bowtie2 short-read alignment tool. However, the second alignment to the target marker sequences was done in two different ways, using BLAST and using Bowtie, respectively.

The breakdown of the read counts corresponding to the above workflow is shown in the following table.

<table>
<tr><th>Raw reads</th><th>Quality control</th><th>Alignment to reference genome</th><th>Alignment to markers using Bowtie2</th><th>Alignment to markers using BLAST</th></tr>
<tr><td rowspan="4">30,709,595</td><td>Bad: 489,892 (1.6%)</td><td></td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td rowspan="3">Good: 30,219,703 (98.40 %)</td><td>Aligned: 27,246,903 (90.16 %)</td></tr>
<tr><td rowspan="2">Unaligned: 2,972,800 (9.84 %)</td><td>Aligned: 122, 938 (4.13 %)</td><td>Aligned: 132,707 (4.46 %)</td></tr>
<tr><td>Unaligned: 2,849,862 (95.87 %)</td><td>Unaligned: 2,840,093 (95.54%)</td></tr>
</table>

We identified the markers using the aligned reads from the second alignment stage. These are the reads unaligned to the reference genome and aligned to the marker sequences, which are 122,938 reads and 132,707 reads, respectively. The identification results from both alignment approaches are shown below.

Markers identified in reference-unaligned sequences in Sample 5

Identified marker		**Bowtie analysis** (Reads whose Bowtie best hit is this marker)		**BLAST analysis** (Reads whose MEGAN assignment using BLAST output is this organism/marker)	
Accession number	**Description**	**Number of reads**	**% reference-unaligned reads**	**Number of reads**	**% reference-unaligned reads**
gi|325965616|gb|JF420885.1|	Lux single copy cloning vector pMH30 luxCDABE operon, complete sequence; and aminoglycoside 3'-phosphotransferase (kanR) gene, complete cds	122,470	4.1197	118,809	3.9965
gi|15430752|gb|AF403784.1|	Photorhabdus luminescens lux operon,complete sequence	426	0.0143	11,356	0.3820
AY183453	Escherichia coli plasmid pIP1100 class II integron multidrug resistance locus, partial sequence.	24	0.0008	32	0.0011
gi|386685405|dbj|AB665981.1|	Staphylococcus aureus DNA, ACME and type II SCCmec, complete and partial sequence, strain: SR 388	6	0.0002	14	0.0005
AB113580	Klebsiella pneumoniae integron In112 genes (intI1,blaGES-3, aacA1, orfG, orfA, qacEdelta1), IS26 tnpA, IS6100 tnpA, orf6, orf5 genes, complete and partial cds.	5	0.0002	8	0.0003
gi|31321897|gb|AY037297.1|	Synthetic construct erythromycin resistance protein (erm) gene, partial cds; and streptomycin 3'-phosphotransferase (sph), bleomycin phosphotransferase (ble), neomycin phosphotransferase (nptII), and gentamycin resistance protein (aac) genes, complete cds	3	0.0001		
JX023441	Acinetobacter baumannii class I integron IntI1 (intI1) gene, partial cds, GES-22 (blaGES-22), aminoglycoside acetyltransferase (6') type I (aac(6')-Ib), and dihydrofolate reductase type A7 (dfrA7) genes, complete cds, and qacEdelta1 (qacEdelta1) gene, partial cds.	1	0.0000		
U12441.2	Escherichia coli plasmid R388 class 1 integron In3,complete sequence.	1	0.0000		
X12868	Plasmid pLMO20 dhfrV gene for	1	0.0000		

	trimethoprim resistance.				
gi|439981485|gb|JX976326.1|	Escherichia coli plasmid pECDF16 extended spectrum beta-lactamase 2 (TEM-1) gene, complete cds	1	0.0000		

The species-level taxonomy tree from the MEGAN assignment of the query sequences using the BLAST output is shown below.